OBSERVATION TRAINING AND PRACTICE: EFFECTS ON PERCEPTION OF BEHAVIOR CHANGE

by

MARK ROBERT WEINROTT

A DISSERTATION

Presented to the Department of Psychology
and the Graduate Faculty of McGill University
in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

October 1975

© MARK ROBERT WEINROTT 1976

MARK WEINROTT, PH.D
PSYCHOLOGY

OBSERVATION TRAINING: EFFECTS ON PERCEPTION OF BEHAVIOR

Abstract

This study examined the relationship between behavior change in children and teacher perception of that change. It was hypothesized that the extent to which an individual is trained in observation skills, practices them, and is monitored by others is related to the accuracy of his ratings of child behavior.

A laboratory test of this hypothesis (Experiment I) showed that teachers who were trained to record discrete responses and collected data on a daily basis were quite accurate in their judgments of distractibility. Ratings by teachers who received little or no training without practice were considerably less accurate.

A naturalistic test of the same hypothesis was also performed (Experiment II). Teacher ratings of a selected child were compared with independently obtained observation data. Results showed that the effects of observation training, data collection, and monitoring were not significant in improving the accuracy of perception.

An in-service teacher training program in behavior modification was also evaluated.

Résumé

Cette étude a examiné le rapport entre le changement du comportement chez les enfants, et la perception de ce changement par le professeur. L'hypothèse était que le niveau d'entraînement d'un individu dans la pratique des techniques d'observation, de son utilisation de ces techniques et de la surveillance par d'autres, est en relation avec la précision de son évaluation du comportement de l'enfant.

Cette hypothèse, testée en laboratoire (première expérience) a démontré que les professeurs entraînés à enregistrer les réactions isolées, et qui recueillaient les données quotidiennement, étaient très exacts dans leur jugement du degré de distraction. Les évaluations faites par des professeurs n'ayant reçu que peu ou aucun entraînement, et sans expérience pratique, étaient considérablement moins exactes.

Un test de la même hypothèse a aussi été effectué dans le milieu naturel (deuxième expérience). Les évaluations des professeurs pour un enfant choisi au préalable ont été comparées avec des données obtenues indépendamment. Les résultats ont démontré que les effets de l'entraînement à l'observation, l'accumulation des données et la surveillance n'étaient pas significatifs dans l'amélioration de la justesse de perception.

Un programme d'entraînement pour la modification du comportement, à l'intention des enseignants, a aussi été évalué.

ACKNOWLEDGEMENTS

"Pulling off" a study of this size requires an extraordinary amount of luck and the cooperation of many people. A log was kept throughout the project and it reveals the names of no fewer than 150 persons who provided assistance. Were they gathered together it would be difficult to determine their common ground--although it would be a hell of a party. They include Jack Hearn of Cadbury Chocolate, Ltd. and Lina DiCicco, whose love notes etched in salami provided an original contribution to art, if not to science. A number of people deserve special thanks.

To John Corson, whose willingness to provide moral support goes far beyond specific contributions to this dissertation. His blind faith in a depressed or otherwise arrogant student was primarily responsible for my remaining in psychology. Never has there been a more clear-cut application of noncontingent attention. For this I will always be indebted.

To Dick Jones, methodologist extraordinaire, who provided many hours of consultation, and who permitted me to see him struggle with problems. This came as a great relief and has since enabled me to persevere about statistical issues without breaking into a cold sweat.

To the freewheeling Social Learning Project and the Oregon Research Institute, whose members showed me that working long hours is really weird. May we continue to be productive for many years to come.

To Marc Wilchesky, whose performance on this project indicates a bright future as a file clerk. His contribution to the administration of the project, to the training of observers, and to my sanity cannot be overestimated.

To Janet Anderson, companion and graphic artist all in one. Her plans and desires were often subverted by this project. I am truly grateful for her ability to divert my attention toward the real world. She is responsible for my surfacing as a human being of sorts.

To Bill Scott, David Hiatt, and Dick Dempster, who were awakened at sunrise every weekday morning for five months; who took hundreds of phone messages, and who were amazingly tolerant of their roommate. May you revel in flank steak forever. Also, a special thank-you to David's Capri, which, despite two accidents and a black spot in Consumer Reports, sputtered and lurched for 5,000 miles in service to this project.

To the team of observers who labored with uncommon dedication for a mere pittance. You were a delightful group to work with and a ravenous bunch of eaters.

To the programmers and analysts, Ron Siegel, Bernie Loftus, Sadru Teja, Bernie Corrigan, and Brian Bauske, who nursed the data through numerous conversions and analyses. Your qualified reassurance was exceedingly therapeutic. Computing assistance was provided by the McGill University Computing Centre, J & P Coates (Canada), Ltd., International Computers (Canada), Ltd., Oregon Research Institute Computer Center, University of Oregon Computing Center, and Health Sciences Computing Facility at UCLA.

To Mary Taylor for editing and typing the drafts and final version of this dissertation. Her ability to decipher hieroglyphics make her an archeological prodigy.

To the Molson Foundation and the Grant Foundation for their generous support of this research.

Finally, to the Channel 6 weatherman for the mild Montreal winter of 1973-1974. The schools were closed only once because of snow.

TABLE OF CONTENTS

Table of Contents (continued)

Table of Contents (continued)

List of Tables

List of Tables (continued)

List of Tables (continued)

List of Figures

Overview

In psychology, as in the physical sciences, a degree of agreement has been attained as to various rules and procedures for observing and reporting events objectively, for minimizing personal bias, and for optimizing reliability in the process of prediction. These generally accepted guidelines and rules fall at numerous points along a continuum ranging from great generality to extreme specificity. One pole relates to the philosophy of science, including a consideration of the overall functions, maxims, presuppositions, strengths and limitations of a discipline; the opposite pole relates to highly specific prescriptions and procedures for utilization of specialized methods of observation and assessment, such as psychometric tests, questionnaires, rating scales, interviews, phenomenological reports and electro-physiological monitoring.

At each point along this spectrum, two questions are relevant. First, given a purpose, a research question or series of questions, what manner of investigative exercises, operations and tactics should one embark upon to fulfill the purpose and answer the questions? Here one is dealing with decision rules concerning the appropriateness or adequacy of a research strategy, particularly with elements of experimental design, selection of independent variables, general specification of dependent measures, and choice of theoretical models to aid in interpretation. The second question, and the one more commonly associated with "methodology" is: How does one obtain interpretable data for which the ambiguity of evaluation is reduced to the lowest possible degree? Here the focus is on clear delineation of dependent variables and procedures for procuring, scoring, and analyzing data. It is to this second question that the present study is addressed.

This investigation examines the relationship between behavior change in children and adult perception of that change. More specifically, it tests the effects of systematic observation on one's perception of a behavioral disorder. It is hypothesized that the extent to which an individual is trained in observation skills, practices them, and is monitored by others, will be related to the accuracy of his perception or assessment of the child's behavior. In accordance with methodological objectives, the purpose of the study is to learn whether adult perception can be rendered less ambiguous and more reliable.

The dissertation is organized in the following fashion: first, a theoretical discussion highlighting the need for behavioral data; second, a presentation of theoretical issues and empirical evidence supporting the use of naturalistic observation; third, a description of Experiment I, an analogue test of the effects of observation training; fourth, a presentation of Experiment II, a naturalistic test for the same effects; and fifth, the evaluation of an ongoing clinical intervention. The dissertation will conclude with a sixth section on the implications for subsequent clinical and research applications of the findings, as well as a statement of the theoretical significance of this work.

The Importance of Behavioral Data

In the early 1960's, when behavior modification began making inroads in clinical psychology, it was discovered that traditional assessment instruments designed to provide information about attitudes, traits, and underlying dynamics simply did not satisfy the requirements of those who advocated a behavioral model. As early as 1934, La Piere demonstrated that attitudes and behavior have no obvious relationship to each other. This incongruence was difficult to

reconcile at an intuitive level, for as Cohen (1964) states:

> Most of the investigators whose work we have examined made the broad psychological assumption that since attitudes are evaluative predispositions, they have consequences for the way people act toward others, for the programs they actually undertake, and for the manner in which they carry them out. Thus, attitudes are always seen as precursors of behavior, as determinants of how a person will actually behave in his daily affairs. [Pp. 137-138]

Wicker (1969) reviewed 33 studies which examined the relationship between attitudes and overt behavior. His overall conclusion was that little evidence is available to support the existence of underlying constructs within an individual which influence both his verbal expressions and his actions. Festinger (1964) and Vroom (1964) rendered the identical conclusion in shorter reviews, while Cohen (1964) was even more skeptical in his suggestion that attitude change procedures do nothing more than cause cognitive realignments, and perhaps that the concept of attitude has no critical significance whatever for psychology.

When a social label (e.g., "deviant" or normal) or clinical diagnosis (e.g., passive-aggressive) is ascribed to an individual, it is presumed that such a construct also mirrors one's actual behavior in some way. However, the observer's or rater's abstractions may be related only tenuously to the response patterns of the subject in question. Research on person perception strongly suggests that the way a person is described depends far more on the observer than on the person observed (Crow, 1957; Hjelle, 1968; Vernon, 1964). For clinical purposes, it seems essential to know if an individual who is labeled "deviant" actually displays more abnormal behavior than one who is described as normal.

A number of studies have found that children who are labeled as deviant and

referred for psychotherapy differ significantly from non-referred children in terms of parent ratings of their traits or behaviors (Conners, 1970; Miller, Hampe, Barrett, & Noble, 1971; Schectman, 1970; Sines, Paulkner, Sines, & Owens, 1969; Speer, 1971; Wolff, 1967). Nevertheless, in-home observations of these children yield minimal differences, or typically, no differences in rates of deviant behavior between referred and non-referred children (Hendriks, 1972; Lobitz & Johnson, 1974; Patterson, Cobb, & Ray, 1972; Shaw, 1971). Research conducted in educational settings shows somewhat greater convergences between (teacher) ratings and observed behavior (Bolstad, 1974; Patterson, Cobb, & Ray, 1972; Werry & Quay, 1968); however, only in comparisons of average students and those who have been identified as extremely deviant have strong behavioral differences between groups confirmed teachers' descriptions.

It seems clear that referral for treatment is based on many factors other than observed rates of noxious responses (Buckle & Lebovici, 1960; Lapouse & Monk, 1958; Rutter & Graham, 1965; Shaw, 1971; Shepperd, Oppenheim, & Mitchell, 1966). Therefore, in planning an intervention, it is useful to determine the degree of relationship between multiple measures and to identify the specific behaviors that account for a diagnostic label or trait in cases where the expected convergence is obtained. This is particularly important when behaviorally oriented treatment is recommended.

On both theoretical (Bandura, 1969) and empirical (Paul, 1969a, b) grounds, it has been shown that total reliance on conventional trait assessment yields data of negligible predictive validity in cases where behaviorally oriented treatment has been implemented. That is, response patterns or attitudes may change without concomitant modification of behavior (Walter & Gilmore, 1973; Wright, 1972). It would seem obvious that some form of evaluation used in a

behavioral context should be consistent with empirical goals. This is particularly true if, as Festinger (1964) suggests, attitude change is inherently unstable and will dissipate or remain isolated unless an environmental or behavioral change can be brought about to support and maintain it.

When this precept is violated, and the data base bears little relationship to the level of the problem, serious errors in evaluation may result. As an illustration, let us examine the treatment of conduct disorders in delinquent youth. This is a problem which is amenable to empirical interpretation, as the criteria for diagnosis involve overt activities of a criminal nature and subsequent adjudication. Treatment outcome measures have always included recidivism or re-entry into the juvenile justice system--a behavioral index which professionals of all theoretical persuasions deem most significant. Halleck (1967), a noted psychoanalyst, admits that:

> The psychiatrist has few more important functions in criminology than evaluating the probability that a given offender is likely to do violence to his fellow man. [p. 313]

Historically, those who have instituted programs for delinquents have relied heavily on non-behavioral assessment devices. In many cases, these have revealed substantial personality adjustment which was assumed to predict subsequent behavior outside the treatment setting (although such devices were not specifically designed for this purpose). The relationship between performance on conventional measures and recidivism is a sad chronicle on the effectiveness of the juvenile rehabilitation system.

Aichorn (1935) pioneered the application of Freudian theory to treatment of aggressive delinquent boys in Vienna during the 1920's. His model of delinquency focused on over-protective parents and the boys' receiving either excessive or inadequate amounts of parental love. Resulting aggressive behavior was

attributable to the interplay of psychic forces engaged in the conflict. The development of positive transference between boys and therapist was viewed as a realization on the part of the boy that adults were caring and trustworthy. On the basis of subjective impression, the primary dependent variable, psychoanalytic treatment of delinquency was a resounding success and its adoption by American clinicians quickly followed.

In the most publicized of these psychoanalytically based efforts, Redl and Wineman (1951) altered Aichorn's original interpretation to include elements of ego weakness and lack of impulse control. The "delinquent ego" referred to a psychic organization which operated in opposition to normally accepted cultural values. Treatment was geared to the development of the super-ego in such a fashion that impulse gratification would be channeled toward more acceptable alternatives. Diminution of major symptomatology (e.g., stealing, vandalism) was considered relevant, but was clearly not the major thrust of treatment. The criteria upon which a boy was considered adjusted included ability to relate meaningfully to image symbols, to use verbal modes of communication, to be less suspicious of adults, and to perceive the necessity for rules and routines. Here again, the problem of a conduct disorder is viewed in dynamic terms with neither treatment nor evaluation consistent with symptomatology. As a consequence, Redl and Wineman (1952) were faced with admitting failure when recidivism in their sample remained high. Despite earlier claims of success, they concluded that:

> ...our 'children who hate' went back into the limbo of the 'children that nobody wants.' This spectacle of their retraumatization of strengths that had been so painfully, if incompletely, implanted in their personalities being literally wasted in a battle in a hostile environment, is one that fades slowly, if at all, from our minds. [p. 315]

This dilemma was not confined solely to psychoanalytic approaches. Weeks' (1958) selection of psychometric tests, self-ratings, and sociometric descriptions may have played a role in establishing the well-known Highfields project as a successful group model for treatment of juvenile offenders. Unfortunately, no follow-up data were obtained. Two additional studies involving social work intervention failed to demonstrate a treatment effect on any behavioral measure (e.g., completion of school, grades, deportment), despite the fact that certain attitudinal or personality indices suggested otherwise (Meyer, Borgatta, & Jones, 1965; Vasey, 1968).

The Cambridge-Somerville Project (Powers & Witmer, 1951) serves as another example of what is likely to occur when problems defined as environmental or educational in nature become reinterpreted as psychiatric. The authors' original conclusion that "none of the evaluative methods employed indicates any degree of success for the treatment program" (p. xix) has been revised to suggest that boys who received psychotherapy yielded a greater likelihood of subsequent arrest and conviction (Cross, 1964; Teuber & Powers, 1953). Again, the data obtained during treatment fail to predict re-entry into the juvenile justice system. This unfortunate state of affairs in the assessment of delinquency was highlighted by Eysenck (1952), whose criticism of psychotherapeutic outcome with delinquents was harsh and not entirely accurate. There has been at least one instance of moderate success using intensive counseling. Adams (1961) showed a greater incidence of favorable discharge from state custody for treated individuals who were deemed amenable according to "pooled clinical judgments." However, when amenable and non-amenable samples were combined, there were no differences between the treated group and untreated controls. The preliminary

results of the behaviorally oriented Teaching-Family model (Achievement Place) lend further support to the position that delinquency is a behavioral problem requiring assessment procedures of an empirical nature (Fixsen, Phillips, & Wolf, 1972; Phillips, 1968).

Juvenile delinquency has been a convenient whipping boy for critics of conventional psychotherapy. Nevertheless, evidence is accumulating which indicates that a number of other disorders may better fit a social learning model than a psychodynamic or sociological one. Among these are aggression (Patterson, 1975), withdrawal (Walker & Hops, 1973), alcoholism (Sobell, Sobell, & Christelman, 1972), obesity (Stuart & Davis, 1972), and depression (Lewinsohn, 1972). In each case, there is an increasing effort to design and implement empirically based measures which serve both as diagnostic instruments and dependent variables.

The discussion thus far has focused on the issues of construct and predictive validity of psychometric or impressionistic data applied to problems that are essentially behavioral. Once the decision has been made to include empirical indices, one must confront the second methodological question raised earlier: How does one generate interpretable data for which the ambiguity of assessment is reduced to the lowest possible degree?

Naturalistic Observation: Theoretical and Empirical Rationale

Naturalistic observation has, historically, contributed little to the systematic exploration of human behavior. The advent of behavior modification, with its empirical foundation and focus on social learning has been largely responsible for the recent adoption of this method by psychologists whose theoretical persuasion is not chiefly ethological. In fact, it has been noted that the most significant contribution of behavior modification may be its reliance upon and refinement of naturalistic observation procedures (Johnson & Bolstad, 1973).

Prior to 1960, mental health professionals rarely examined psychological disorders within the context of currently prevailing environmental factors. Indeed, the preferred method of research often obscured whatever pattern and organization may have existed within the natural environment. Experimental treatments were applied exclusively in institutions, clinics, or practitioners' offices; diagnostic and outcome measures consisted of structured personality inventories, projective tests, questionnaires, rating scales, verbal self-report, and therapists' impressions. Total reliance on these measures has been sharply criticized (Webb, Campbell, Schwartz, & Sechrest, 1966):

> Today, the dominant mass of social science research is based upon interviews and questionnaires. We lament this over-dependence upon a single, fallible method. Interviews and questionnaires intrude as a foreign element into the social setting they would describe, they create as well as measure attitudes, they elicit atypical roles and responses, they are limited to those who are accessible and will cooperate, and the responses obtained are produced in part by dimensions of individual differences irrelevant to the topic at hand. [p. 1]

Because such measures revealed little about the influence of the social and physical environment on the organization of behavior, the result was more

intensive study of, and dependence upon, the "black box," a quasi-empirical, theory-determined attempt to ascribe performance to hypothetical constructs such as traits, attitudes, or needs. Testing for functional relationships within the social environment was seldom considered a viable alternative to the "black box." And, in those few instances where it was, there were available no systematic means of analyzing and interpreting naturalistic data. For example, as recently as 1955, 18 full day records of children's behavior were collected by having observers dictate reports of all events as they occurred (Barker, Wright, Barker, & Schoggen, 1961). While such efforts to study naturally occurring phenomena are to be commended on theoretical grounds, logistical problems accompanied by numerous sources of possible artifact[1] prevented the rapid emergence of a strong movement in this direction. In addition to the fact that only a small amount of usable data could be extracted from the massive volume of transcripts, the following problems have been cited as interfering with efficient collection and analysis of naturalistic data: use of participant (vs. independent) observers, difficulty in obtaining control groups, lack of a system for encoding or reducing complex interactions to interpretable units, ethical considerations (e.g., invasion of privacy) and, above all, failure to couch hypotheses in terms of overt behavior (Boyd & DeVault, 1966; Willems & Raush, 1969).

About 10-15 years ago, operant researchers began systematic gathering of naturalistic data in a coded or abbreviated fashion. This was carried out in the home (Wahler, 1969), school (Harris, Wolf, & Baer, 1964), and institution (Ayllon & Azrin, 1965). Assignment of codes was accomplished by breaking down

---

[1] E.g., the reliability of the observers was not assessed.

relevant global patterns of behavior (e.g., aggression) into their component responses (i.e., hit, yell, tease, destruction). Clear operational definitions for each coded behavior were formulated and memorized by independent (non-participant) observers so that subjective judgment was minimized. Hence, recording could be done rapidly.

Hereafter, use of the term "naturalistic observation" refers to (a) the recording of behavioral events at the time they occur, (b) the use of independent, trained observer coders, and (c) descriptive responses which require a minimum of inference to be coded (Jones, Reid, & Patterson, 1975). These criteria preclude parent or teacher report data of any type, regardless of their compliance with rules a. and c. Furthermore, the term does not include global ratings or reports by independent assessors, as these are necessarily retrospective and rely extensively on subjective impression.

Suppose one assumes that a given observation (or rating) x is subject to various sources of error (e) which, if inoperative, would enable x to represent a totally accurate or "true" observation. In deference to the hazards of engaging in a philosophical consideration of "truth," the author proposes the following definition for reasons of mathematical utility. A "true" score shall be one in which corresponding observed and error scores are uncorrelated and in which error scores on different administrations are also uncorrelated (Lord & Novick, 1968).

While the relative magnitude of "e" for both ratings and coded observations would be virtually impossible to ascertain outside of the laboratory, it is possible to identify the major sources of potential error which differentially affect the two types of scores in field settings. Four source dimensions are relevant: simultaneous vs. retrospective collection, molar vs. molecular units,

global vs. specific reporting, and participant vs. non-participant observation. All of the studies that have examined the degree of association between ratings of child behavior and naturalistic observation allow at least two of the source dimensions to vary concurrently. Global reports are, by nature, retrospective, as are most behavior checklists. Discrete responses (molecular units) are often recorded by participant observers. Consequently, it is difficult, if not impossible, to determine which experiments are most relevant to a particular dimension. For purposes of organization, an attempt has been made to arbitrarily cite various studies in which at least part of the measurement error is due to a particular source.

It seems clear that retrospective data are highly selective and reflect characteristics of the observer more than attributes of a subject (Vernon, 1964). Nevertheless, as indices of perception, retrospective ratings may represent social values. Therefore, they may be generalizable to a variety of culturally relevant criterion situations. This is particularly true since criterion measures generally involve subjective judgments, opinions, and ratings of "significant" others (Wiggins, 1973). In other words, retrospective measures may be more valid externally than internally. Such a position supports the use of these measures for specific purposes.

Notwithstanding, inherent measurement error may be substantial. In a series of interviews, it was discovered that mothers' current perception of earlier experiences and attitudes often showed little relationship to similar ratings taken at the time of those events (Haggard, Brekstad, & Skard, 1960). Schnelle (1974) failed to find the slightest relationship between parents' written estimates of prior school attendance and the actual attendance pattern, despite the fact that such behavior is clearly defined and easily monitored.

The accumulation of data on molecular behavioral units or discrete responses (e.g., vacant staring) within a behavior class (e.g., social withdrawal) is a strategy which typically yields higher levels of inter-observer agreement than do conventional forms of trait attribution (Becker, 1960; Walter & Gilmore, 1973). This may be accounted for by the use of descriptive categories in naturalistic observation rather than omnibus categories that employ evaluative judgment. Reliability of evaluative measures requires agreement on both the topography and the intent of a behavior. At a more global level, it depends upon consensus as to the "value" of an attribute. Mischel (1968) points out that the use of summary reports for trait assessment is based on a variety of cognitive and perceptual processes producing "constructed consistencies" in evaluation. When attribute levels (i.e., global ratings) remain stable despite fluctuations in their manifestations (behaviors), traits can be seen as constructs of the observer rather than as attributes of the subject. A totally different level of human judgment (with less susceptibility to the observer bias) is involved in the ongoing recording of discrete behavior units and reporting of data in terms of amplitude, frequency, rate, and duration of specific responses. Proponents of this form of collection contend that the division of a given attribute (e.g., withdrawal) into a number of narrowly defined components and the extensive sampling of these components will yield a result of greater generalizability than a more global rating.

It should not, therefore, be assumed that retrospective behavior ratings are accurate reflections of real behavior (Novick, Rosenfeld, Block, & Davidson, 1966; Wiggins, 1973; Yarrow, Campbell, & Burton, 1964). Indeed, the evidence demonstrates only a very weak relationship at best. Adult ratings of child deviance and observed deviance have produced low-level, generally non-significant

correlations (Guerney, Shapiro, & Stover, 1968; Lobitz & Johnson, 1974).

The tendency in most clinical studies has been for parents to overestimate treatment effects.[2] Clement and Milne (1973) concluded that parental reports of improvement seemed unreasonably favorable in comparison to less reactive measures including observed behavior frequencies. In the same study, parents in a no-treatment control group reported improvement to the same degree as parents in the treatment conditions, despite clear differences on other forms of assessment. Collins (1966) showed that parents reported significant improvement in their child's behavior even though therapy had not yet begun. Walter & Gilmore (1973) found that parents in a placebo-control (pseudotherapy) group reported positive changes in behavior. Further, their rated expectancies for progress remained high, despite the fact that home observations indicated that child behavior was becoming increasingly deviant. Mothers' and fathers' descriptions of child behavior have shown only low or moderate correlation (Eron, Banta, Walder, & Laulicht, 1961), while the relationship between parent and teacher ratings is even lower (Becker, 1960). Investigators have failed to show a consistent relationship between child symptoms reported globally by parents and rates of noxious behavior observed by independent observers in the home (Hendriks, 1972; Tharp, Wetzel, & Thorne, 1968) and in the laboratory (Honig, Tannenbaum, & Caldwell, 1968; Radke-Yarrow, 1963; Sears, 1965).

Poor estimation of child behavior levels is not confined to parents. Bernal (personal communication) found non-significant rank order correlations

---

[2] Several hypotheses regarding the over-estimation have been offered. For example, parental ratings may represent one's conception of an ideal relationship (Becker, 1960), a tendency to portray the family as a cultural stereotype (McCord & McCord, 1961), or recent exposure to recommendations of a reknowned child-rearing expert (Robbins, 1963).

in comparing observation data with a behavior checklist completed by teachers. Moreover, the relative proportions of students assigned to deviant and normal groups by the observation data were significantly different from the checklist classification. A number of other studies have demonstrated that teacher ratings of pupil behavior do not show significant convergent validity with observation data (Jones & Cobb, 1973; Krumboltz & Goodwin, 1966). Wickman (1928) and Maccoby and Masters (1970) have also raised doubts as the validity of teacher ratings, while Wahler and Leske (1973) have demonstrated that teachers' summary reports were not indicative of actual levels or changes in child behavior.

Bolstad (1974) found that scores on a measure of teacher attitude and a behavior checklist did not correlate significantly with each other or with observed rates of appropriate or off-task behavior. Furthermore, proportion scores of attending behavior were significantly positively correlated with reading achievement. It has also been demonstrated that specific academic behaviors such as attending to task, talking to teacher about academic material, volunteering, and talking to peers about academic material were significantly related to achievement scores in reading (Cobb, 1970) and arithmetic (Cobb, 1972). More recent research (Cobb & Hops, 1973; Hops & Cobb, 1973; Walker & Hops, 1974) has demonstrated a functional relationship between specifically taught facilitative behaviors and achievement. It has also been shown that classroom behavior predicts academic achievement over the school year about as well as intelligence tests and that the addition of behavioral information to test scores provides a more accurate prediction of achievement than that obtained by either measure alone (McKinney, Mason, Perkerson, & Clifford, 1975).

Although the majority of studies have shown that parents and teachers are unreliable observers even when tracking a single, well-defined class of behavior, there is contradictory evidence which supports convergence between this form of monitoring and independently observed behavior. It has been suggested that if parental reports for relatively discrete categories were limited to the preceding 24 hours, these would indeed correlate with actual frequencies (Douglas et al., 1968). Using a modified version of this strategy, Peine (1970) did, in fact, obtain data yielding high intra-subject reliabilities. Nevertheless, frequency data were grossly in error. Noncompliance was underestimated by as much as 700%. In another, aforementioned study (Walter & Gilmore, 1973), attention-placebo control subjects gave global ratings which were in glaring contrast to the overall pattern of deviant behavior. Yet, they were able to track on-the-spot specific target behaviors with accuracy. It is interesting to note that this practice failed to influence global assessment. Hines (1974) found that adults combined accurate personal observations with an experimenter's diagnostic label in forming an overall impression of a child. This label ("deviant" or "normal") carried the heaviest weight in this combination. When the information from the two sources of data was contradictory, the induced expectancy determined adults' ratings of the child. Patterson, Shaw, and Ebner (1969) also assumed that parents could collect accurate data when category definitions were precise. Later research failed to support this conclusion as the parents' daily report showed no relationship whatever to observation data (Patterson, personal communication). On the basis of this finding, Patterson and his co-workers have implemented a daily telephone interview procedure in which only the occurrence or non-occurrence of selected behaviors is noted (Jones, 1974).

Despite the fact that both participant and non-participant observation may produce reactive effects, it again appears that coding as practiced by an unbiased, independent observer has the advantage of reducing the error variance. Harris (1969) showed that mothers' data were not as reliable as those obtained by non-participant observers. Also, the participant or familiar observer has the potential for exerting greater control over the events to be recorded than a less active or familiar counterpart (Mash & Hedley, 1974). Herbert and Baer (1972) found that the mean percent agreement between mother and observer coding mother's attention to appropriate behavior was 46%. When two independent observers were used, the reliability rose to 90%.

It would appear that teacher or parent ratings are susceptible to more sources of error than simultaneous recording and reporting of molecular units by independent observers. As such, the latter form is considered to be more representative of the "true score" as defined earlier; thus it can be said that ongoing recording of discrete behavioral units by an independent observer is more objective. It would seem, on that basis alone, that one could safely employ them as standards against which to compare less reliable forms.

In an effort to train individuals to become more accurate observers in their own setting, Wahler and Leske (1973) conducted an experiment in which 40 elementary school teachers viewed a series of 15 video tapes depicting six children engaged in independent seat work. The children were actually following prepared scripts which systematically determined the percentage of time they were working appropriately or behaving in a distractible fashion. One of the children's distractible behaviors was faded over the 15 segments, such that she produced off-task behavior on 75% of the first tape, 70% of the second, and so on, with each subsequent tape portraying a decrease in off-task responses of

5%. One group of teachers was given instruction in how to sequentially sample behavioral events and record frequencies of specified responses, while another group received no such input. At the conclusion of each tape, all teachers were required to rate each child on a seven-point "distractibility" scale. Results showed that the untrained teachers were quite inaccurate in their ratings of the target child, most failing to report even slight "improvement" until the 13th tape (when the child was only 25% distractible, or after a 50% shift). The trained teachers were considerably more accurate in their appraisals, as ratings of most were sensitive to a shift of only 15%).

In a successful replication of this study, Leslie (1975) examined the direction of the behavior change as well as the effects of systematic observation. Subjects viewed a series of five video tapes in either gradually "improving" or "deteriorating" order. Compared to subjects who passively observed, those who applied the prescribed tracking techniques perceived a greater amount of change and estimated significantly less overall deviant behavior.

Neither of these analogue studies presents protocols in a sequence which represents a non-linear pattern of behavior. While improvement may be relatively gradual, it may eventually reach an asymptote and occasionally revert toward baseline. Reversion is particularly likely when contingencies are withdrawn abruptly. The extent to which improvement continues to be reported during a plateau phase and that to which subsequent deterioration goes undetected are important criteria in the assessment of observation skills.

Taken as a whole, these studies offer little comfort to the behavioral clinician or researcher whose assessment data are generated solely through means of interview, questionnaire, or psychometric test battery.

Wahler and Leske (1973) state, with ample justification, that parents and teachers will continue to provide reports that summarize a cluster of discrete responses (e.g., immaturity, aggressiveness). Behaviorally oriented interventions typically include pinpointing and tracking overt target behaviors (e.g., noncompliance, out of seat), but it is not uncommon for both therapist and observer to lapse into more global descriptive statements even when more specific data are available. One tactic has been to discourage the use of constructs. However, in a field which clings to relatively global psychiatric jargon, it might be more productive to improve the accuracy of observation and thereby improve both the use of this jargon and the resulting summary reports.

In an attempt to replicate the findings of Wahler and Leske (1973) and Leslie (1975), the present study used similar procedures in a modified form. It was deemed desirable to separate the effects of the initial observation training itself from those of practicing the skills on a routine basis. Certainly, it is not uncommon for adults to receive training in tracking followed only by encouragement for employing the skill. Whether systematic data collection is carried out depends largely on the whims of the individual. Many clients are simply not convinced of the value of complying with the suggestion. Consequently, this newly acquired skill may well be lost after a period of inactivity.

One problem which arises when individuals are, in fact, tracking behavior is a decline in their accuracy over time. In a study by O'Leary and Kent (1973), fixed groups of trained observers who were restricted to observation and computation of reliability coefficients within their own membership began to drift in their application of a behavioral code. Although intra-group reliability (agreement) remained high, recordings on pre-coded video tapes

gradually showed less relationship to the actual pattern and to data obtained from other groups of observers. This phenomenon was attributed to the development of idiosyncratic definitions of behaviors to be coded. There seems to be a strong possibility that a single trained observer who receives no feedback throughout the course of weeks or months may also be coding according to gradually shifting criteria. For instance, in order for "disapproval" to be coded during the baseline phase, an observer might require that a child display some form of a subtle tonal quality in addition to a negative verbal statement. However, a few weeks later, the same statement might be recorded as "disapproval" in the absence of this tone. The observer may feel that he is adhering to the original definition when he is, in fact, drifting. In order to prevent this from occurring, researchers have rotated observers such that reliability checks are made on all possible pairs. In addition, periodic retraining on pre-coded tapes has been instituted (Johnson & Bolstad, 1973). While these precautions are feasible for use with trained professional observers, they are much less appropriate for teachers or parents who may be expected to collect data over an extended period of time. Still, their tendency to drift might well affect both their recording of specific behaviors and also any summary reports they might submit. One possible means of removing this artifactual variable is to use a rotating, calibrating observer or external monitor who would be responsible for recording simultaneously, on a time sampling basis, the same behaviors designated for the parent or teacher. For general clinical use in the school setting, the use of parent aides, student teachers, secretaries, or various administrative personnel could serve this purpose. The teacher would define to each external monitor the topography of the response required for inclusion in a particular code category. Following a period in which both

were recording individual behaviors for the same subject(s), a comparison of results could be made. In this manner, a teacher might be prevented from drifting as she would be calibrating her reliability with a number of supplementary observers. It is hypothesized that such a procedure will improve the accuracy of both frequency and duration data, and hence, the quality of summary reports.

Hypothesis: Experiment I

Teachers' ratings of child behavior will show greater convergence with independently observed levels of distractibility as a function of training, practice, and feedback in systematic observation.

Method: Experiment I

Subjects

Ss were 40 elementary school teachers (grades 1-3) enrolled in a continuing education course in behavior modification at McGill University. Participants were matched on certain target child variables (to be described in Experiment II) and were then ranked and consecutively assigned to one of four groups (n = 10) as follows:

E_1: Ss received no training in observation skills or data gathering techniques and were not encouraged to attempt systematic assessment of any kind. In order to control for differences in subject-instructor contact hours, E_1 received a two-hour placebo input on pharmacological intervention with hyperactive children in lieu of the observation training session.

E_2: This group received one two-hour session of observation training consisting of instruction and practice in systematic viewing of classroom interaction and collecting of data on specified behaviors (e.g., noncompliance, out-of-seat). Using video tapes, teachers learned three common recording procedures: event recording, which provides measures of frequency of occurrence of target behaviors; duration recording, which provides measures of the duration of occurrence; and occurrence-nonoccurrence (interval) recording, which can provide estimates of both the frequency and rate of the target response. Where appropriate, the obtained data were converted into rate per minute, or proportion of intervals in which the behavior had occurred. The session involved practice in deciding which sampling procedure was most efficient yet would still yield a valid representation of the behavioral levels in question. This was followed by application of the selected strategy. A minimum of six target behaviors and accompanying tapes served as practice material. Teachers were given encouragement to use the techniques in their own setting, but were not required to do so.

E_3: Members of this group received one two-hour session of observation training identical to that described above, plus the assignment of collecting data daily ("tracking") on selected behaviors (2-3) emitted by a target child in their own classroom. In addition, teachers were required to submit a weekly

written record of the data to the instructor in order to gain admission to the session. (This group is analogous to Wahler and Leske's experimental group; the daily data collection is comparable to test trials conducted on consecutive days).

E_4: Ss received one session of observation training and were assigned the task of daily data collection as described for group E_3. It was further stipulated that members of E_4 would recruit a third party monitor (e.g., student teacher, parent aide, free-flow teacher, assistant principal) from within the school who would observe the target child and record data along with the teacher. This form of "reliability" assessment was to be carried out for a minimum of three 15-minute periods per week. Each teacher was responsible for training her own calibrators using predetermined definitions and observation strategy. The identity of the external monitor changed periodically but not systematically, as a partial control against observer drift in a fixed teacher-monitor pair. Admittance to the weekly course meeting was contingent upon the teacher submitting both sets of data to the instructor.[3]

---

[3] Teachers in all four conditions were accustomed to having aides, student teachers, etc., on a regular basis, so that E_4 does not differ along a dimension of adult contact.

Procedure

One week prior to observation training (or placebo input for E_1), all Ss were presented with a strategy for pinpointing and defining target behaviors. Each teacher had isolated three target behaviors for a preselected child in her class and had been required to formulate definitions for these. Two weeks following the observation training (or placebo), teachers viewed the first of a series of seven video protocols, each of which depicted two boys seated at adjacent desks whose rates of off-task (distractible) behavior were systematically manipulated. The behavior of one child was varied such that he emitted distractible behavior during 70% of the first tape, 60% of the second, and so on until he was off-task on only 30% of the fifth protocol. On the basis of episodes 1-5, this child's behavior could be construed as "improving." A matched set of 30% and 40% off-task tapes was also produced and constituted trials six and seven respectively. The behavior of the other boy was held relatively constant at a level of 35-45% off-task. Children were assigned the task of independently solving arithmetic problems presented in workbook.

The protocols were produced by directing the children to follow prepared scripts which prescribed the topography of the response each would emit. The 10-minute segment was divided into 40 15-second intervals. At the beginning of an interval, individual instructions were given to the children by a director who wrote them on a blackboard. The boys were asked to produce one or two of the following responses: out-of-seat, talking with peer, manipulate object, look around, or work. In cases where a child was told to produce two responses within an interval, these were to occur sequentially, not concurrently, and always involved shifting from one distractible behavior to another. When a

task-relevant response was evoked, it was carried on for the entire 15 seconds. The responses were randomly assigned to intervals within the script, although the ratio of appropriate to inappropriate behavior was pre-arranged for each child.

Teachers observed one tape per week for seven weeks. They were told that the protocols were to be used for training independent observers on a classroom coding system, and that the purpose of screening them was to assess the complexity[4] of the tape by determining the level of distractible behavior. Tapes used for training were to be catalogued in this manner so that the instructor could better evaluate the performance of the independent observers. Ss were not given an instructional set with respect to an expected pattern of child behavior, nor were they advised of any diagnostic labels.

Just before showing each tape, the instructor wrote the four distractible response categories on the blackboard and asked the teachers to refer to this list in any way which would help them make a more accurate overall appraisal. Members of the three experimental groups were asked to record frequencies for each. Teachers were instructed to watch each tape carefully, to remain silent, and at the conclusion, to place a mark on a seven-point "distractibility" scale at the point which best described how the target child compared to students with whom the teachers ordinarily dealt (i.e., peer norms). Teachers retained no record of their ratings from week to week.

---

[4] Number of different behavior categories required to accurately code the sequence.

Results: Experiment I

Groups and standard deviations for ratings on each protocol are presented in Table 1.1. The data were analyzed by assigning to each of the seven protocols an ideal or standard rating to which obtained scores could be compared. Because the amount of distractible behavior decreased in a linear fashion (trials 1-5) it was assumed that totally accurate ratings should depict a similar pattern. Ratings on tapes six and seven (30% and 40% distractible) should coincide directly with those for protocols five and four, respectively.

Table 1.1.

Group Means and Standard Deviations for Ratings

Protocol	E_1	E_2	E_3	E_4
Means				
1	6.4	6.6	6.3	6.8
2	6.2	6.4	5.5	6.1
3	6.4	6.5	5.0	5.5
4	5.3	5.5	4.0	4.5
5	3.6	4.1	3.2	3.5
6	4.2	2.7	2.9	3.0
7	3.8	5.5	3.4	3.3
Standard Deviations				
1	.84327	.69921	.48305	.42164
2	.91894	.69921	.70711	.73786
3	.69921	.84984	1.24722	.70711
4	1.49443	1.26930	.66667	.84984
5	1.07497	1.52388	1.47573	1.08012
6	1.87380	.94868	.87559	.94281
7	1.31656	1.17851	1.50554	.67495

The ideal rating for the first protocol was identified as "7." Each tape (through no. 5) was assigned a rating one point lower than the previous tape.

Virtually any standard could have been used as long as the order and magnitude of the differences between protocols was preserved. The assignment of number "7" to the first tape offered the added advantage of representing the modal and median ratings for each of the four groups of subjects. Table 1.2 presents the standard ratings for each protocol, while Table 1.3 shows the means and standard deviations for deviation scores.

Table 1.2

Standard Ratings for Each Protocol

	Protocol number						
	1	2	3	4	5	6	7
Percent of distractible behavior	70	60	50	40	30	30	40
Standard rating	7	6	5	4	3	3	4

Figure 1.1 shows the mean group ratings for each trial and the corresponding standard ratings. A repeated measures ANOVA was performed on the seven deviation scores for the four groups. Results are presented in Table 1.4. Main effects for groups ($F = 7.014$, $df = 3, 36$; $p < .001$) and protocols ($F = 4.625$, $df = 6, 216$; $p < .001$) were found, as well as a significant interaction between the two factors ($F = 1.804$, $df = 18, 216$; $p < .03$). Orthogonal comparisons (Winer, 1971) between the four group means were performed, revealing differences between the two pairs of groups. E_1 and E_2 deviated from standard ratings significantly more than did E_3 and E_4, the members of which were collecting daily data ($F = 17.643$, $df = 1, 180$; $p < .01$). Thus, the null hypothesis may be rejected. Group E_2, which received one session of observation

training, did not differ from controls. However, those groups conducting daily data collection demonstrated greater sensitivity to changing proportions of distractible behavior. The addition of an external monitor was not a significant factor in improving the accuracy of E_4 although its overall mean deviation score of .586 was lower than that of .843 for E_3 ($F = 3.773$, $df = 1, 36$; $p < .1$).

Table 1.3

Means and Standard Deviations for Deviation Scores

Protocol	E_1	E_2	E_3	E_4
Means				
1	.6	.4	.7	.2
2	.8	.6	.7	.5
3	1.4	1.5	1.0	.7
4	1.7	1.7	.4	.7
5	.8	1.3	1.2	.7
6	1.6	.7	.7	.6
7	1.0	1.7	1.2	.7
Standard Deviations				
1	.84327	.69921	.48305	.42164
2	.42164	.51640	.48305	.52705
3	.69921	.84984	.66667	.48305
4	.94868	.94868	.51640	.67495
5	.91894	1.33749	.78881	.94868
6	1.50554	.67495	.48305	.69921
7	.81650	.82327	1.03279	.67495

It was considered important to determine at which point teachers in each group initially perceived "improvement." A series of chi-square analyses were performed to test for independence of groups on the basis of the number of subjects whose ratings had dropped by a minimum of one point from the first

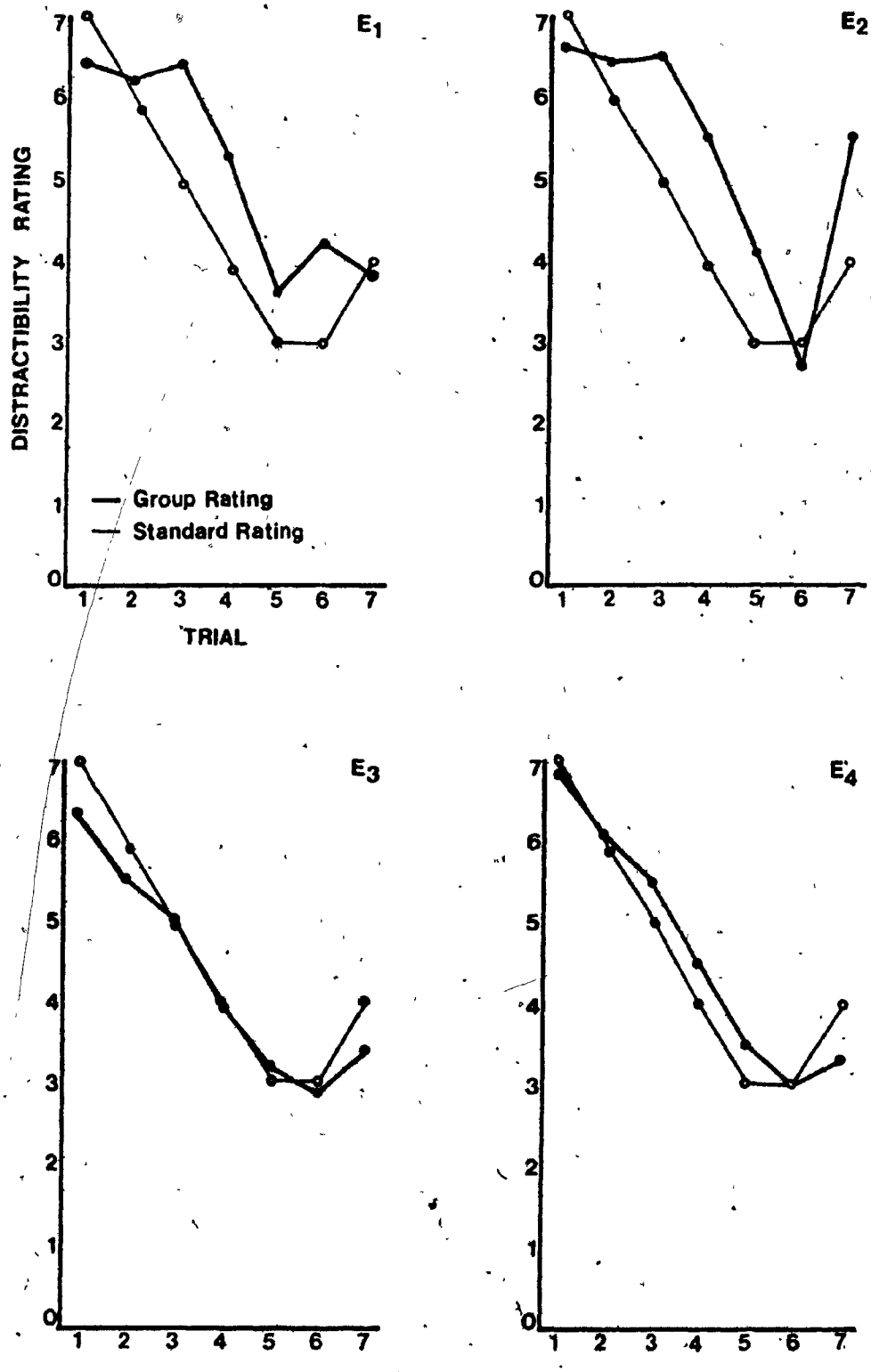


Figure 1.1. Obtained vs. Standard Ratings for Each Protocol

trial. While 14 of 20 teachers in E_3 and E_4 detected the 10% change, this was true for only five members of the remaining 20 Ss (see Figure 1.2). Group differences were found ($X^2 = 9.123$, $df = 3$; $p < .05$) and an *a posteriori* test of independence showed that E_3 and E_4 were superior to E_1 and E_2 in perceiving improvement ($X^2 = 8.120$, $df = 3$; $p < .05$). Group differences ($X^2 = 17.143$, $df = 3$; $p < .01$) and differences between combined groups ($X^2 = 16.942$, $df = 3$; $p < .01$) were still apparent when the criterion for detection of improvement was a minimum one point drop by the third (50% distractible) protocol. Seventeen of the 20 teachers tracking daily lowered their ratings; only four of 20 in the other groups did. At the point where the target child was off task only 40% of the time, five of the 20 teachers in E_1 and E_2 still had not responded to the change, while all members of E_3 and E_4 had recorded at least a one-point decrease; this difference was not statistically significant for groups ($X^2 = 6.171$, $df = 3$; n.s.) or combined groups ($X^2 = 5.714$, $df = 3$; n.s.). It would appear that observation training combined with daily tracking enabled the majority of the teachers to detect a 10% increase in task-oriented behavior. Most of those who did not receive both training and practice required changes of 30% before their summary reports were altered (see Figure 1.1).

A test of the ability to maintain a constant rating on the second (matched) 30% protocol (trial 6) revealed no differences ($X^2 = 3.29$, $df = 3$; n.s.). Many subjects in each experimental group, and all in E_2, demonstrated a "halo" effect (i.e., they perceived less distractible behavior). Only the control group mean was higher on the second 30% trial than the first ($\Delta\bar{x} = +.6$, a rather surprising finding, given that an instructional set or expectation for continued improvement had allegedly been developed by experimental subjects. Despite the disparate trend demonstrated by E_1, the only group difference

appeared between E_1 and E_2 (Newman-Keuls, $p < .01$), the latter yielding a mean of -1.4 between ratings on matched protocols. This indicated a substantial "halo" effect. Difference scores between trials 5 and 6 for E_2 and E_3 were -.3 and -.5, respectively.

Table 1.4

Analysis of Variance of Deviation Scores

Source	SS	DF	MS	F
Mean	237.72707	1	237.72707	349.11157
A (Group)	14.32831	3	4.77610	7.01390***
Error	24.51414	36	.68095	
B (Protocol)	17.02135	6	2.83689	4.62520***
A X B	19.92108	18	1.10673	1.80438*
Error	132.48470	216	.61336	

* $p < .05$
** $p < .01$
*** $p < .001$

Somewhat different results emerged between ratings on the sixth and seventh protocols (i.e., mild deterioration). E_2, following a demonstration of the most extreme "halo" effect observed, differed from all other groups in that all 10 of its subjects reported heightened distractibility ($X^2 = 14.164$, $df = 3$; $p < .01$). Only two Ss in group E_1 reported deterioration, while seven Ss in E_3 and five in E_4 detected a change in the proper direction. An ANOVA was performed to test for differences in magnitude between a subject's rating on the second 30% tape plus one. It was assumed that a one-point increase was commensurate with the degree of actual regression in the target child's on-task behavior. A main effect for groups was found ($F = 3.045$, $df = 3, 36$; $p < .05$). Orthogonal comparisons showed that E_1 and E_2 were, once again, more

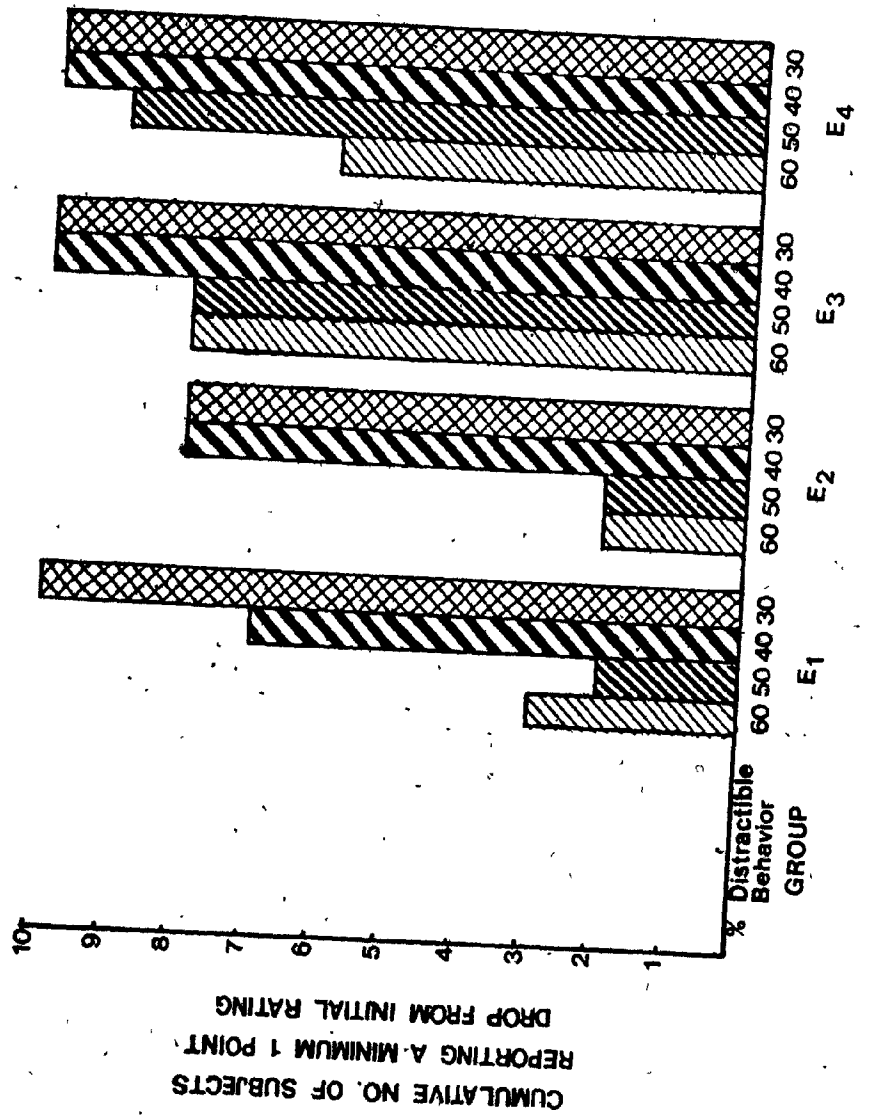


Figure 1.2. Cumulative Number of Ss in Each Group which Reported a Minimum One-Point Drop in Distractibility from Their Rating on Trial 1

variable in their ratings, hence, less reliable than observers in E_3 and E_4 ($\underline{F} = 8.955$, $\underline{df} = 1, 36$; $\underline{p} < .01$).

Discussion: Experiment I

The results of this study strongly support the hypothesis that subjects who systematically observed and tracked child behavior on a daily basis perceived changes in behavior levels earlier and with greater accuracy than those who did not collect daily data. Most interesting was the discovery that a "one shot" training session was ineffective in improving the accuracy of summary reports. A weekly restatement of behavioral definitions and encouragement for collecting frequency data during test trials was not sufficient to raise the performance of group E_2 over that of untrained control subjects.

It may be that a logical rationale and periodic encouragement are simply not powerful enough to raise motivational levels so that observers will conscientiously apply relevant techniques. This suggests that the impetus for accurate tracking is generated by the tracking itself. In other words, the systematic observation is perceived as useful only after it has been employed. Such an interpretation coincides with the cognitive dissonance model (Festinger, 1957), whereby effort may be perceived as warranted simply because it has been exerted. Another possibility is that the daily data provide a form of feedback to the observer which is viewed as both interesting and potentially useful. Such a phenomenon is often reported by newly trained observers and this could serve to heighten motivation for subsequent tracking.

The effects of the training itself cannot be isolated from motivational components in the present study. Whether E_2 failed to retain the skills or lacked the incentive to apply them diligently in test trials is a matter of conjecture. Members of this group did, indeed, record frequencies which were characterized by moderately higher variability than those obtained by E_3 or E_4. This suggests greater individual differences in definitions, attending,

or recording for E_2. Since all members of this group detected an increase in distractible behavior on the final trial, it is unlikely that a failure to attend to the stimulus accounted for a major portion of the variance in frequency data. The fact that ratings for E_2 did not show a great deal more variability in early trials than those obtained for E_3 and E_4 suggests that summary reports were filed without consideration of recorded frequencies. Such a tendency has been noted by a number of investigators (Johnson & Bolstad, 1973; O'Leary & Kent, 1973; Scott, Burton, & Yarrow, 1967; Walter & Gilmore, 1973). It may be that the subjects who were collecting daily data in their own classrooms had come to rely on these as a basis upon which to file global reports.

The present study was a successful replication of experimental effects demonstrated by Wahler and Leske (1973). Despite differences in design and procedures that would mitigate against obtaining similar results in the present study, the findings were virtually identical. These factors included fewer stimulus children to be observed (2 vs. 6), the rating of only one child (vs. 6), and larger decrements in distractible behavior (10% vs. 5%). Each of these should have contributed to a more easily detected change; on this point, it is worth nothing that nearly all subjects reported improvement at the point where distractible behavior occurred 40% of the time compared to 25% of the intervals in the Wahler and Leske experiment.[5]

Also corroborated were several findings reported by Leslie (1975), who found that systematic observation functioned to improve the accuracy of

---

[5] Differences in inter-trial intervals (one week vs. one day); the length of protocols (10 vs. 15 minutes) and the latency between training and initial test trial (two weeks vs. several minutes) may have served to enhance the probability of successful replication.

perception regardless of whether child behavior was improving or deteriorating. Results of the present investigation confirmed that both the proportion of individuals detecting early improvement and the overall deviation of ratings from standard criterion levels were sensitive to observation training and practice. When the stimulus conditions changed such that the target child displayed a higher degree of distractibility, there was considerably less divergence from the expected ratings for groups E_3 and E_4. The fact that all members of E_2 reported deterioration on the final trial may have been due to the abnormally large "halo" effect observed on the previous trial. While all three experimental groups perceived some improvement from one 30% distractible protocol to a matched tape, the magnitude of the mean difference was approximately three times greater for E_2. Having lowered their ratings substantially when behavior levels remained stable may have enabled a contrast or compensatory effect to occur when the deterioration was detected. In other words, the final two ratings may not have been independent of each other. A parsimonious interpretation would be that there existed no obvious differences between the number of individuals in each of the three experimental groups who detected deterioration. However, the proportion of control subjects (E_1) reporting a shift in the expected direction was considerably lower than that of their experimental counterparts. In fact, the mean ratings for the control group on the final trial were actually below that for the previous one despite the increase in distractibility. It is also surprising that this group demonstrated no discernable "halo" effect; distractibility was actually rated higher in the second 30% trial than the first.

As the overall variability for E_1 and E_2 was higher than that for the daily tracking groups, the members of the tracking groups were more reliable,

or objective, in the sense that observer agreement with a standard criterion constitutes a measure of reliability as well as accuracy.

Superiority of the two tracking groups is probably accounted for in three ways. (1) The process of systematic observation requires continuous activity in identifying behavior, and making rapid, subtle discriminations on the basis of well-formulated definitions. (2) A product of this process is a personal record of observed frequencies or durations. (3) The process requires the careful distribution of attention to the target child in the teacher's own class.

The discovery that one session of observation training was insufficient to raise observational skills above the level obtained by naive observers was not predicted. It would appear that following this instruction with tracking assignments is a necessary condition for increasing the accuracy of summary reports. The addition of an external monitor did not appreciably enhance the quality of reports submitted, although the overall deviation from criterion standards was lowest under this condition. Perhaps the reason the monitor contributed relatively little to daily tracking could be attributed to the brief duration of the study (seven weeks of test trials) during which observers may have remained well anchored to their definitions of behavior categories. Another possibility is that daily data collection did not always focus on the same behaviors targeted in the test protocols. Consequently, external monitors would have served little purpose other than to insure that tracking was indeed occurring. Regular use of systematic observation techniques has (potentially) important implications for (a) clinical treatment, (b) the quality of interaction between adult and child, and (c) for field research in program evaluation.

It is a well-known fact that children experiencing academic and behavioral problems can be identified early in their school careers (Cobb, 1970; Robbins, 1966; Walker, 1971). Failure to do so is likely to have far-reaching consequences for a child. It has been shown that nearly half of the high school students who are academically deficient display a "spread pattern" from elementary through secondary school. That is, they were initially failing in only one or two subject areas but gradually deteriorated in an increasing number of others (Fitzsimmons, Cheever, Leonard, & Macunovich, 1969). A parallel can be drawn to children experiencing behavioral difficulties where stability of disturbance appears very strong (Walker, 1971; Zax, Cowan, Rappaport, Beach, & Laird, 1968). Many educators recommend remedial programs during the early elementary grades with particular emphasis placed on basic academic skills and those facilitating behaviors requisite for learning. Such programs are often expensive, overextended and free of evaluations. Given that benefits do accrue, it is important to recognize these so that remedial work can procede efficiently or be eliminated altogether. Similarly, the benefits of identifying an ineffective intervention for a particular child are considerable. Systematic observation and routine tracking could serve to enhance the identification and periodic assessment procedures already employed.

On the basis of the present experiment, one could argue that selective and, perhaps, exclusive, attention to negative behaviors could account for a failure to report improvement. However, this is not likely. Leslie (1975) found no differences between summary reports of obsrvers who recorded positive behavior and those who focused on deviant responses. Similarly, it was noted that trained observers expressed greater "likability" for the

stimulus child than did naive observers, regardless of the valence or directionality of behavior recorded. It would appear that the use of observation training will render a teacher's (or parent's) evaluation more objective and his attitude toward the child increasingly positive. Modification of adult attitude toward a deviant child is considered by some to be the most important effect of any treatment program and the most reliable predictor of outcome (Eyberg & Johnson, 1974).

With the current emphasis on program evaluation and accountability, the importance of scientific precision is increasing. That results have often been misinterpreted or ignored by consumers and policy makers is both an invitation to and reflection of compromises in the processes of subject selection, instrumentation, data collection, analysis, interpretation, and dissemination. Such compromises render any scientific demonstration an approximation of the facts. It is becoming obvious that one cannot afford the price of approximation in data collection when reallocation of resources necessitates reductions in other areas. To the extent that low-cost systematic tracking produces more reliable data in field settings, it should be utilized, irrespective of other abuses in the evaluation process.

It must be re-emphasized that the evidence of a "tracking effect" has been derived largely from analogue studies which may not be representative of naturally occurring phenomena. The relationship between a teacher or parent and a problem child carries with it a number of affective or motivational properties which may differentially influence the quality of, and reliance upon, behavioral data. Observer bias, demand characteristics, evaluation anticipation, and various setting variables are sure to differ from laboratory to classroom or home. It is not known whether the aspects

which vary are those functionally relevant to the "tracking effect," or whether the degree of difference is sufficient to produce discrepant results between settings. The generalizability of the "tracking effect" is an empirical question to be addressed in Experiment II.

Introduction: Experiment II

Repeated attempts to relate perception and behavioral change have resulted in two sets of results deriving from an equal number of methodological approaches. The clinical literature is relatively consistent in showing that adults who are untrained in observation skills tend to over-estimate the effectiveness of counseling (Teuber & Powers, 1953), psychotherapy (Paul, 1966), psychoanalysis (Lazarus & Davison, 1971; Redl & Wineman, 1952), behavior therapy (Eyberg & Johnson, 1974; Wright, 1972), and pseudotherapy (Walter & Gilmore, 1973) when behavioral indices are used as criterion variables. Yet, naive observers in the analogue studies cited previously (Leslie, 1975; Wahler & Leske, 1973) displayed the opposite effect. That is, improvement had to be very substantial before it was perceived by untrained observers.

The crucial element lacking in the analogue studies may well be an instructional set or expectancy for an emerging pattern of child behavior. Viewing a sequence of video vignettes in a laboratory setting (these tapes depicting children with whom the observer shares no experience or mutual affective involvement) is clearly different from the typical mode of observation and evaluation. A teacher who initiates a referral, seeks training, and carries out a treatment program with her own student is subject to environmental influences which may bias her impression of the child. These effects may have little relation to variables that tend to influence her evaluation in a contrived situation.

A second difference between the two types of studies concerns the complexity of the behavior patterns observed. Despite the inclusion of three Conditions in Experiment I (gradual improvement, stability, and mild

deterioration), behavioral data typically show high variability from hour to hour (Johnson, Christianson, & Bellamy, 1974), day to day (Jones, 1972), and activity to activity (Walker, Hops, Greenwood, & Todd, 1975), a characteristic not represented in the video procedure. Conversion of accurate data on discrete responses to a summary report covering a number of observations is a more arbitrary and difficult task than that required in Experiment I. In the naturalistic case, one is expected to disregard the "noise" in the system and synthesize a set of frequencies, rates, and durations into a global appraisal. These data may be tainted by subjective considerations of, or allowances for, extraordinary circumstances, including variation in schedules, seating arrangements, materials, changes in behavior of adjacent peers, weather, or an anticipated activity. Each adds an element of variability or "noise" into the classroom which an observer need not contend with in rating a film-mediated stimulus child. It may be that the vast quantity of "noise" causes an overload on a teacher's perceptual system to the extent that it is simply not capable of processing the input. Rather than disregard extraneous variables, they are considered supplementary to frequencies, durations, or proportions that have been obtained and become incorporated into a summary report which may bear little relationship to the response levels themselves. To an extent, this had been noted for group E_2 in Experiment I. The tendency to disregard "hard" data if it fails to corroborate a global evaluation or Gestalt may be even greater in the natural setting.

A third factor which could interfere with accurate global appraisal in the classroom setting concerns a possible dependency between behavioral evaluation and academic achievement. Though not yet subjected to empirical investigation, it seems reasonable to hypothesize that children of high

academic standards, showing equivalent rates of noxious behavior as academically deficient children, may be rated as less deviant (behaviorally). Since no academic information was provided about stimulus children in analogue studies, this dependency would be operative only in the actual classroom.

Finally, in assessing the performance or behavioral level of a child in one's own class, a teacher is aware of both the purpose of the evaluation and the implications of scores for subsequent academic or behavioral programming. To the extent that she believes a particular course of action should be taken, the results of the assessment may be inadvertently biased. Certainly, the literature on experimenter bias (Rosenthal, 1969) and reliability of therapist inferences (Scott, Burton, & Yarrow, 1967) lends support to this notion. The analogue studies contain no such element of future consequences and, as such, may be construed as removing artifact which operates in the real situation.

There are numerous reasons why one would not expect the demonstration of a "tracking effect" in the natural environment. Yet, on the basis of unequivocal laboratory results, it is hypothesized that training and practice in observation techniques tends to make an individual's summary reports converge with observed behavior levels in both the laboratory and the natural environment, regardless of the directionality of the error (i.e., the tendency to either underestimate or overestimate response levels). The present experiment attempted to replicate Experiment I, but in a number of actual classrooms.

A secondary objective of the present experiment was to test the effectiveness of an in-service teacher training program in behavior modification. Curricula have been developed by Andrews (1970), Becker, Engelman, and Thomas

(1971), and Hall (1971) and are becoming increasingly common. Unfortunately, the effects of indirect intervention by means of teacher training have not been carefully documented. There appears to have been little attempt to evaluate such programs using multiple measures which include a sufficient number of naturalistic observations. Seldom have matched control groups been employed and long-term follow-up is more the exception than the rule. One characteristic of most programs is a preoccupation with management of noxious responses. Techniques for the modification of behaviors such as fighting, arguing, out-of-seat, and non-compliance are widely espoused, almost to the exclusion of intervention strategies for behaviors displayed by phobic, immature, anxious, or socially withdrawn children (O'Leary & O'Leary, 1972; Patterson, Cobb, & Ray, 1972).

The present study followed what could be described as a standard sequence of behavioral inputs but placed added emphasis on the treatment of social withdrawal. The rationale for including this population was twofold. First, accuracy of teacher perceptions may vary as a function of presenting problem. Children described as "socially withdrawn" are characterized by lower than normal rates of behavior in areas of interaction and assertiveness, while the garden variety acting-out child can be thought of as displaying behavioral excess in these areas. Observing the acquisition of skills requiring new topographies may be much different from tracking changes in rates of behavior which already exist to a moderate degree (e.g., attending).

Further justification for developing and evaluating treatment procedures for social withdrawal lies in the historically sparse consideration afforded this problem. These children are seldom, if ever, disruptive, and are not characterized by serious academic deficits. Consequently, their condition is

not normally treated with urgency by either clinicians or scientific investigators. It is, in fact, characterized by a repertoire of behaviors which divert attention rather than attract it. Socially withdrawn children appear to follow four rules for purposes of preserving their anonymity in school: (1) never to be caught out in the open; if possible, don't be seen at all; (2) always keep a line of students between you and the teacher's eye; (3) use the vacant eyeball ploy when cover is thin or unavailable; (4) stay away from all peers who are big, loud, popular, verbal, cute, or otherwise conspicuous.

To combat the frequent use of these strategies, several behavioral techniques have been adopted. An exhaustive list includes adult social reinforcement (Allen et al., 1964; Milby, 1970), adult social reinforcement plus priming (Baer & Wolf, 1970; Buell, Stoddard, Harris, & Baer, 1968; Hart et al., 1968), modeling with guided participation (Ross, Ross, & Evans, 1971), symbolic modeling (O'Connor, 1969) symbolic modeling plus shaping (O'Connor, 1972), stimulus fading (Conrad, Deck, & Williams, 1974), desensitization plus shaping (Reid et al., 1967), social reinforcement plus tangible rewards (Calhoun & Koenig, 1973; Kale & Toler, 1970; Whitman, Mecurio & Caponigri, 1970), group and individual contingencies (Walker & Hops, 1973; Walker, Hops, Greenwood, & Todd, 1975). These studies are dominated by single case reports of treatment applied directly by professionals. The total absence of matched control groups is rather surprising since it is often stated, with little documentation, that socially withdrawn children will "grow out of it" if just given time.

The present intervention examined the degree to which a prepared sequence of teacher-administered treatment procedures would be effective in increasing

levels of peer interaction, volunteering and initiating to teacher, and in reducing the frequency of daydreaming (looking around) and self-stimulation. This treatment "package," along with one appropriate for use with disruptive-distractible pupils, was evaluated. Both were implemented as part of an in-service teacher training program which precluded direct professional contact with identified problem children.[6]

Evaluation was undertaken using somewhat different standards than ordinarily employed in operant research. It is customary to compare behavior levels in intervention, termination, and follow-up to those observed in baseline. Such baseline observations are conducted in as unobtrusive a manner as possible, generally without the imposition of an instructional set on the subjects. This does not, however, guarantee that subjects are behaving naturally. Baseline measures do not necessarily depict adult or child performance under conditions of high motivation. It may well be that given an instructional set, adults may prove to be effective in maintaining more desirable behavior over the course of an observation than would be represented in a "natural" baseline. Perhaps they have at hand some of the resources which can modify behavior, but do not normally utilize them.

The present study introduced a "demand baseline" procedure in which teachers were asked to try to increase the social and assertive behavior of socially withdrawn children or to decrease the noxious responses of acting-out children. Whatever means were employed for this purpose were left to the imagination of the individual teacher. The decision to introduce a demand

---

[6] It is important to note that the single session of observation training described in Experiment I was only a small part of the behavior modification program.

baseline was based on recent evidence pertaining to the operation of demand characteristics and observer presence effects.

It has been suggested that demand characteristics operate most potently during intervention and follow-up probes when clients feel they are expected to perform certain operations which were not part of their repertoire during baseline. Rosenthal (1969) offers a number of possible explanations for this phenomenon, including desire to please the experimenter (therapist), anticipation of evaluation, and the possibility that the observer may become a discriminative stimulus for certain kinds of interaction patterns.

The extent to which demand characteristics operate differentially in baseline and intervention, both in magnitude and direction, may account for a sizable portion of variance normally attributed to innovative treatment techniques. It may be that individuals being considered for treatment have an interest in "making" their children "look bad" in order to justify their request for assistance. Johnson and Lobitz (1974) and Lobitz and Johnson (1974) found that parents had the ability to manipulate their children's behavior on request. This holds true for both normal and deviant families, although the latter group is less maleable toward the positive. There is every reason to expect that teachers have the same capacity, particularly since they more closely represent "normal" parents than their deviant counterparts.

Baseline data are vulnerable to instructional set or expectations, and they may also be sensitive to observer presence. In a well designed study, Kent, Fisher, and O'Leary (1974) discovered an interaction between observer presence and phase of treatment (baseline vs. intervention) using child deviant behavior as the dependent variable. School children displayed higher

rates of noxious responses in baseline when observed overtly, but lower rates during intervention. Rates of deviant behavior during treatment under a covert (via one-way glass) observation condition were actually higher (22%) than in baseline. The "improvement" observed using overt monitoring would ordinarily be attributed to a main treatment effect instead of to reactivity. Unfortunately, Kent did not gather data on teacher behavior during baseline so it is perhaps premature to attribute any shift in child behavior to fluctuations in the teacher's own response pattern. Nevertheless, it seems clear that demand characteristics operate differentially in conditions of covert and overt monitoring. Taken together, the three studies cited suggest that baseline may be artificially inflated or depressed due to a number of variables generally considered as artifact.

Johnson and Lobitz (1974) offer two recommendations for allaying the impact of demand characteristics: first, to make observations less intrusive, and second, to rely on multiple measures in testing hypotheses pertaining to treatment effects. Efforts have been made to reduce the conspicuousness of observers (e.g., prebaseline observations, restricted interaction between observer and students). Yet, there are serious logistical and ethical limitations to the widespread use of covert observation. The value of employing multiple dependent measures cannot be over-emphasized, particularly since adult attitude may, in some cases, be an accurate predictor of treatment outcome (Eyberg and Johnson, 1974). However, this merely circumvents the issue of obtaining observational data of greater convergent validity. Both of the aforementioned suggestions were taken into account in the present study. In addition, the demand baseline was introduced. It is assumed that performance levels observed during intervention which vary from those derived during a high-

demand baseline can be attributed more to treatment procedures themselves than to motivational variables.

The following hypotheses were tested in Experiment II.

Hypothesis I

The accuracy of teacher perceptions of behavior change in children (as defined by convergence with observational data obtained by independent agents) will bear a direct relationship to the amount of observation training, practice, and monitoring which teachers receive.

Hypothesis II

Given only an instructional set, teachers will be capable of manipulating the behavior of selected problem children in a socially desirable direction.

Hypothesis III

Acting-out children whose teachers receive behavior modification training will emit lower levels of total deviant, disruptive, and distractible behaviors following intervention than were displayed in either the natural or demand baseline conditions.

Hypothesis IV

Acting-out children whose teachers receive behavior modification training will emit lower levels of total deviant, disruptive, and distractible behaviors than matched control children whose teachers received no such training.

Hypothesis V

The behavior of acting-out children whose teachers have completed behavior modification training will not differ from same-sex classroom norms for total deviance, disruption, and distractibility.

Hypothesis VI

Socially withdrawn children whose teachers receive training will show a lower level of withdrawn behavior following intervention than was displayed in either the natural or demand baseline conditions.

Hypothesis VII

Socially withdrawn children whose teachers receive training will show a lower proportion of withdrawn behavior following intervention than will untreated control children.

Hypothesis VIII

Socially withdrawn children whose teachers receive behavior modification training will not differ from same-sex classroom norms for withdrawal.

Naturalistic Observations

Observers and Recruiting Procedures

The following advertisement was placed in the "Woman Help Wanted" classified section of The Montreal Star on two consecutive Saturdays:

> Observers for interesting research project in child psychology. Will train, approx. 20 hrs. week, begin mid-December, B.A., married, car absolutely necessary. 842-1241 ext. 1627, Monday to Friday.

One hundred fifty-five inquiries were received by a secretary, who performed an initial screening to insure that all conditions stated in the ad were fulfilled. She offered no further information, and merely recorded the applicant's name, address, and telephone number. A research assistant telephoned each for details of salary, duration, and nature of the position. An attempt was made to emphasize the negative aspects of the job (low wages, irregular schedule, variable hours, extensive traveling, and a five-month commitment to the project). This was done to discourage less motivated candidates. If the candidate was still interested, an appointment for an interview was arranged. Fifty-one individual interviews were conducted by E in order to describe the observation coding system and to ascertain whether the applicant appeared sufficiently intelligent, organized, and personable. The interview was intentionally held in a relatively obscure room in a building which was difficult to locate, thus simulating one component of school observation. All applicants who arrived late for their appointment were rejected, regardless of their performance in the interview (which was merely a formality at that point).

While several investigators have administered a battery of aptitude tests in order to select those easiest to train and potentially most accurate (Kent, personal communication; Skindrud, 1972b), there is no substantive evidence

confirming the predictive validity of the instruments employed. In fact, anecdotal reports indicate that there is no obvious relationship between observer reliability and test scores (Kent, personal communication; Patterson, personal communication). Consequently, no such measures were administered, thereby reducing selection criteria to subjective impressions of the interviewer. In deference to the likelihood of several poor choices, 10 candidates were hired with the understanding that the two with the lowest reliabilities during training would be terminated prior to baseline.

All candidates agreed verbally to a commitment of five and one-half months, with hourly wages of $2.00 during training and $2.75 thereafter, reimbursement of travel expenses, a maximum of 15 one-hour observations per week, and a weekly retraining session to be held on each Friday afternoon. Every effort was made to limit travel and conform to observers' personal schedules. However, it was made clear that each would be required to visit every classroom a minimum of two times.

Training

One week prior to the initial training session, each observer was mailed a copy of a coding manual based on a system developed by Patterson, Cobb, and Ray (1972). The manual is shown in Appendix A. Observers were instructed to read the manual and memorize the codes and definitions as they would be tested on these upon arrival at the first meeting.

The 10 trainees were divided into two groups. Ninty-minute sessions were scheduled five days per week for three weeks. These were held in a large, quiet room equipped with a Sony 3650 video tape deck, an Electrohome 23" monitor, and an inexpensive cassette tape recorder used for signalling intervals.

The first, third, sixth, eleventh, and seventeenth sessions began with a written test of the codes and category definitions. These were graded by the observer trainer (E or a research assistant) who returned them the following day for review. Trainees spent the remainder of the first session viewing a low complexity video tape of classroom interaction while the trainer modeled the correct coding of the behavior of one student. In the next session, this procedure was expanded to include responses by the environment (i.e., teacher and peers) to the "target child."

The following 14 sessions (through the end of week three) consisted of practice coding of video taped protocols of increasing complexity. The trainer ordinarily coded along with the observers for periods of 2-10 minutes, after which the sequence was replayed for purposes of feedback. During meetings 12 to 15, reliability checks between all possible pairs of observers within a group were conducted on two 10-minute protocols.

In an effort to ease observers into actual classrooms, sessions 16 and 17 were conducted in a "mock" classroom using 5-7 children and a teacher, all of whom were instructed as to which behaviors to emit during each five-minute vignette. All 10 observers were present and seated at the front of the room. A tape recorder signalled each six-second interval. Following each vignette, observers compared their recordings with one other observer. Observers rotated with each trial so that all possible pairs of observers were tested for reliability. Inter-observer agreement was computed and controversial events were discussed publicly until an appropriate policy decision was made.

Following this routine, observers were scheduled to practice individually (2-3 hours) and in pairs (1-2 hours) in actual classrooms prior to the start of the project. Classes which had been observed only once during the search

for subjects were selected as practice sites so that all subjects were exposed to observers twice in advance of baseline.

Finally, four additional reliability checks were made on pre-coded 10-minute protocols as part of a research project on the effects of over-training on observer reliability (Wilchesky, 1974). Two test trials were run on each of two consecutive days (sessions 18-19), the remainder of the sessions being devoted to administrative details and distribution of materials. The total number of hours devoted to observer training and related activities was approximately 30.

Observers were equipped with a clipboard, audio pacer, optical scanner coding forms, a package of travel directions to each of the schools, a map of the Montreal urban community, forms for obtaining classroom rules from teachers, a directory of all teachers and target children, including addresses and phone numbers, a checklist to fill out each day before leaving home, a set of instructions about how the observation should be conducted, and their first weekly schedule. These materials are shown in Appendix A.

Preparation of Video Tapes

During the first two weeks of training, most of the protocols used were copies of tapes originally produced at the Point of Woods School, State University of New York at Stony Brook. These were taken from a fixed camera position, through one-way glass, and depicted two boys seated at adjacent desks. The tapes were rated in terms of their difficulty to code and had been used for purposes of training observers in a number of studies (O'Leary & Kent, 1973). The teacher appeared infrequently in these protocols which created problems in learning environmental response categories. To mitigate

this deficiency, additional tapes with high rates of teacher-student interaction were produced by E; these new tapes were recorded with interval signals on the original sound-track. A total of 20 hours of observer training tapes was produced for use in the present investigation.

Observer Agreement During Training

The criterion of successful training was a minimum of 70% observer agreement on all categories using interval-by-interval computation. That is,

$$\frac{\text{Number of agreements}}{\text{Number of agreements plus disagreements}}$$

Three 15-minute protocols of equal complexity[7] were used for test purposes during the third week of training. These protocols were more difficult to code than actual classroom interactions for a number of reasons. First, the mean complexity of the protocols was higher than that of most live situations (.49 vs. .41). Second, the percentage of environmental responses to the subject's behavior was much higher on the protocols (36% vs. 19%). Third, the audio quality of the tapes was not as high as that of direct sound. And fourth, the restricted vantage point of the camera is responsible for some confusion about the location of materials, the blackboard, teacher, and peers. Therefore, it was anticipated that observer agreement would be substantially higher under natural conditions.

Mean observer agreement on the three test trials was 72% over all

---

[7] Complexity is defined as the number of unrepeated behaviors (i.e., code categories) required to describe an observation segment divided by the total number of possible categories. Taplin and Reid (1973) found that the correlation between percent observer agreement and complexity of criterion protocols was -.52 ($p < .001$). This indicates a tendency for reliability to drop when observed interaction becomes increasingly complex. Reid (1973) replicated this analysis and found a correlation of -.75 ($p < .01$).

categories. This is approximately the same overall level of agreement attained by O'Leary and Kent (personal communication), the original producers of the test tapes. While the objective for the video taped protocols of 70% on each category was not met, the degree of overall reliability attained in the laboratory was considered sufficient. For both mock classroom trials and practice sessions in actual classrooms, the mean reliability coefficient was 85% (range 92% for "physical aggression" to 77% for "look around").

Retraining

Weekly "booster shots" as suggested by Taplin and Reid (1973) and Johnson and Bolstad (1973) were conducted on Friday afternoons following the start of data collection. During the first two months of observations these sessions began with a test on the codes and category definitions. Following the second month these tests were limited to every other week. Observers were not altogether pleased about repeated administration of the test, particularly since scores for each were often perfect. Rather than eliminate the testing procedure, it was decided to introduce a random monetary contingency whereby a perfect score on selected weeks earned the observer five dollars. "Booster shots" consisted of the trainer giving observers feedback on specific categories for which reliabilities were lower than desired, or on the lowest category if all reliabilites were acceptable. Difficult situations were role-played and discussed until a policy was agreed upon. The remainder of the session was devoted to viewing video protocols and to distribution of the following week's schedule.

Experimental Phases

Twenty one-hour observations were conducted for each of the 50 subjects in Experiment II. Observation times were random, but restricted to available one-hour blocks during which the target child was in the referring teacher's classroom.

Baseline I. Three and one-half observations were conducted under baseline conditions typical of most operant research. Teachers were instructed to carry out their normal program without alteration of any kind.

Demand baseline. At the midway point of observation four and at the beginning of the following two observations, the observer handed the teacher a sealed envelope, the contents of which stated:

> If your target child is socially withdrawn, please try to make him appear as outgoing as possible during this observation without deviating from your normal routine.
>
> If your target child is disruptive, please try to make him appear as quiet and cooperative as possible during this observation without deviating from your normal routine.
>
> The observer is not aware of the content of this instruction.

Observers were given a specified number of sealed envelopes and their weekly schedule dictated when and to whom these should be distributed. Observers had no knowledge of which number observation they were assigned to or the identity of control subjects, who did not participate in the demand baseline, but received an equivalent number of additional observations under an extended natural baseline condition. Teachers in the four experimental groups had advance knowledge that the instruction was forthcoming and were told by E that he wanted to see just how effective they could be under optimum conditions of motivation. He conveyed an expectation of improved child behavior but was

also careful to acknowledge that one's ability to manipulate the behavior of a child for a short time interval might require measures which would be difficult to impose on a regular basis.

Baseline II. Three observations under normal baseline conditions were conducted in order to examine residual effects of the previous manipulation and to allow such effects to dissipate prior to intervention. The nine pre-treatment observations described thus far were conducted in a six-week period.

Intervention. Eight observations were held at a rate of one per week, beginning the second week of treatment.

Follow-up. Three weekly observations were scheduled, beginning two weeks following the completion of treatment.

Considerations in Naturalistic Observation

Observer Reliability

The assessment of observer accuracy is of paramount importance in research involving field observation. This is generally conducted by calculating the agreement between two or more observers, but has also been determined by comparing their coding with precoded or criterion video tapes. The first procedure yields a measure of observer agreement or reliability, while the second represents accuracy. The relationship between the two is more complex than one might expect.

O'Leary and Kent (1973) demonstrated that when observers were divided into separate groups and restricted to computation of reliability within their own membership, they soon began to "drift" in their application of a behavioral code. Observer agreement remained consistently high within the group; however, when compared with precoded video protocols, there was a gradual and significant decline in accuracy.

An investigation by Romanczyk, Kent, Diament, and O'Leary (1973) demonstrated that observers showed an immediate drop in reliability (agreement) following training when they were monitored covertly. Periodic overt spot checks analogous to typical calibrating procedures produced a return to high agreement. However, this was restricted only to the overtly monitored sessions (Reid, 1970). Even when given feedback concerning this decline during overt monitoring, observers were able to maintain high levels of agreement for only one subsequent session before the recurrence of drift (Reid & DeMaster, 1972). A random check procedure, in which observers were told that a percentage of their coding sheets over a number of observations would be checked for

agreement against protocols, produced a higher mean reliability than that obtained for both a no-check and a spot-check group. However, at no time did their performance exceed that of the spot-check group on occasions of overt monitoring (Taplin & Reid, 1973). Unfortunately, field observation usually precludes covert monitoring; hence, a random procedure would be difficult to employ. Similarly, the assessment of accuracy vis-a-vis reliability would be impossible due to the absence of a criterion measure. It appears that researchers will have to be content to continue using the periodic spot-check procedure, whereby agreement is assumed to represent accuracy.

A number of hypotheses have been offered to account for the observer drift phenomenon. Divergence between fixed pairs of observers may be due to the development of idiosyncratic definitions of the behaviors to be recorded (O'Leary & Kent, 1973). Or, when only one calibrating observer is used during the course of a study, other observers may change their coding styles to match that of the calibrator (Romanczyk, Kent, Diament, & O'Leary, 1973). The reactive effects of testing (monitoring) must also be considered as a variable which serves to heighten motivation for accuracy and increased vigilance (Johnson & Bolstad, 1973). Another possibility is that there may develop implied, "private" contracts between pairs of observers which "simplify" the events to be recorded. In other words, if one observer notices a subtle response which she believes her partner (calibrator) has overlooked, she may ignore it and code only the more conspicuous behavior even though she would have included it if she were coding alone. Support for this interpretation is derived from a study by Reid (1973), who found that the complexity of behavior situations was lower during reliability checks than for observations conducted individually. However, observer presence effects heightened by the introduction

of a calibrator might also contribute to a reduction in the complexity of overall behavior.

The implication of these findings is that reliability assessment for the purpose of providing feedback to observers in the field must be based on criteria which remain consistent with original category definitions. This can be accomplished in two ways. First, the preparation of a large set of pre-coded videotapes for use during training and periodically during data collection would serve to anchor observers to a fixed standard. Frequent feedback and retraining sessions may serve to ameliorate the degree of drift in non-monitored observations. This procedure was used by DeMaster (1971); she gave bi-weekly "booster shots" and found that the procedure was moderately effective in maintaining levels of reliability. A second possible procedure is to have all observers calibrate all other observers so that a single calibrator does not inadvertently cause others to conform to her particular style of coding.

In the present study, a number of precautions were taken to mitigate declining reliability. Observers were over-trained, as suggested by Taplin and Reid (1973). Initial training lasted well beyond the suggested three-week period (Patterson, Cobb, & Ray, 1973; Skindrud, 1972b), despite the attainment of satisfactory agreement coefficients within that period. Retraining sessions were held weekly after data collection had begun. A large set of video protocols was available, so that observers did not view the same tape twice. Reliability checks occurred very frequently; every seventh observation during baseline and every fifth during intervention and follow-up (i.e., twice per week per observer). Finally, all observers were used to calibrate all other observers.

Observer Bias

The consideration of observer bias as a possible contaminating variable in naturalistic observation was an offshoot of the experimenter bias research conducted largely by Rosenthal and his associates in the mid-1960's (Rosenthal, 1966, 1969). Despite misgivings about methodology and statistical analyses used in a number of these studies, considerable support has been presented confirming the existence of "experimenter effects or error that is asymetrically distributed relative to the 'correct' or 'true' value" (Rosenthal, 1966).

The now common use of independent assessors arose as a reaction to subjective global impressions which seemed most prone to bias. However, an independent rater or observer is not necessarily objective, though he may be more so than an experimenter, therapist, parent, or teacher (Johnson & Bolstad, 1973; Rapp, 1965; Scott, Burton, & Yarrow, 1967).

Kass and O'Leary (1970) made the first systematic attempt to assess the impact of instructional set on recordings of observers in a simulated field experimental situation. Observers were trained on a nine-category coding system and assigned to groups which differed in terms of information each received about the relationship between teacher reprimands and their presumed effect on disruptive behavior. Despite the fact that each group observed the identical sequence of protocols, their recorded rates of noxious behavior differed. In this study the observers were not highly trained, a factor which may well have accounted for this result.

Skindrud (1972a) failed to corroborate this finding when comparing data obtained by skilled professional observers who were aware of family treatment

status with recordings of two calibrating observers who were not. Despite the fact that the former group was informed of the normal vs. deviant and baseline vs. treatment status, their recordings under both overt and covert monitoring revealed no differences from those obtained by their uninformed counterparts.

The most carefully designed experimental work in the area was contributed by Skindrud (1972b), who manipulated expectations about the effect of father's presence on rates of child inappropriate behavior. Again, no bias was reflected in these data. This finding has been supported by O'Leary and Kent (1973), who concluded that knowledge of predicted results exerted no discernable impact on recorded rates of classroom behavior.

The three studies which failed to detect an observer bias effect used highly trained observers whose reliabilities were carefully monitored throughout the course of the investigation. Still, it has been demonstrated that a combination of observer knowledge of predicted effects and evaluative feedback from an experimenter can produce biases in observational data (O'Leary, Kent, & Kanowitz, 1975). Highly skilled observers were systematically and explicitly reinforced for providing data which conformed to an experimental hypothesis, a situation not likely to exist in field research.

While differential expectations or instructional sets do not necessarily produce observer bias, a number of precautions were taken in the present study to insure that observers remained objective. First, the use of a complex coding system with clearly defined, operational categories; second, continuous, six-second interval recording which prevented extensive interpretation of ongoing activity; third, observers were kept uninformed as to the type and length of experimental phases, the composition of the teacher groups, the existence of

control subjects, and the specific purpose of the research (although they were aware that some form of teacher training was implemented). Observers received no evaluative feedback other than that pertaining to reliability.

On the basis of available evidence and the safeguards taken to insure objectivity, it was assumed that the contribution of observer bias to variance in the present data was negligible.

Observer Presence Effects

Only in recent years have systematic attempts been made to assess the extent to which non-participant observation serves as a social stimulus. In two studies cited earlier (Johnson & Lobitz, 1974; Lobitz & Johnson, 1974), it was confirmed that parents have the ability to alter significantly their children's behavior if they are so inclined. The extent to which this occurs when no explicit demands are made is a key question in research on the reactive nature of the observational process.

Harris (1969) found that when mothers surreptitiously observed their own families and the data obtained were compared to those collected by independent, trained observers, there were no significant differences in recorded rates of social interaction, nor were rates of deviant behavior affected. The only reactive effect seemed to be a heightened variability or lack of predictability in the behavior of family members.

Two studies suggest that activity level when defined as distance traversed is subject to a reduction in the presence of an observer. This seems to apply both to adults (Bechtel, 1967) and children (White, 1972). In the latter investigation, it was also found that older children's deviant behavior is suppressed by observer presence, while younger children's deviancy is seemingly

unaffected by the presence of observers. Such empirical evidence lends support to the intuitive notion that younger children are apparently less self-conscious and more prone to ignore an observer (Barker & Wright, 1955; Baumrind, 1967; Johnson & Bolstad, 1973).[8]

Hagen, Craighead, and Paul (1975) failed to demonstrate reactivity of mental health technicians to the presence of an observer. Neither the rate of staff activity nor the qualitative performance of programmatic interactions was affected by overt or covert monitoring. It is worth nothing that the data being collected were known to be used for evaluative purposes, an element which should have contributed to reactivity.

There are five studies that have examined the dimension of reactivity in the classroom. Gussow (1964) used retrospective data derived from narrative reports in concluding that the observer and the subject were involved in a continuous and developing relationship. However, it is important to note that the data Gussow examined were not generated for purposes of testing for reactivity and that serious methodological flaws render any conclusions based on these data as very tenuous. Masling and Stern (1969) used trained observers who coded the behavior of teachers and pupils in 23 different classrooms. Two days of observation were treated as a number of five-minute units which were analyzed to test the hypothesis of dissipating observer effects over time. The notion that there would be less correlation between the initial units and the later ones than between later units and the last one was not confirmed.

---

[8] White (1973) also reported that deviant behavior increased over time. However, it should be noted that the obtained reactive effects were stronger than they might have been if the order in which observer-present and observer-absent conditions had been counterbalanced.

Either reactivity was minimal or relatively stable over a two-day period. Neither of these two studies included an observer-absent condition, which precludes direct manipulation of the independent variable. Also, the abbreviated experimental periods provide little evidence, pro or con, regarding the commonly held belief that reactivity diminishes over time. Surratt, Ulrich, and Hawkins (1969) found that observer presence increased task-oriented behavior above the level detected using a concealed camera; however, this finding is seriously contaminated by the use of an observer who had previously dispensed tokens to the class, contingent upon appropriate behavior.

In an attempt to alleviate both the deficiencies of indirect measurement and too few trials, Mercatoris and Craighead (1974) used a video camera deception procedure to gather data in a single classroom over a 30-day period. The teacher was led to believe that the camera was operative only when a live observer was present. It was, in fact, functioning during both conditions (observer present and absent) of an ABAB design. Randomly coded video tapes indicated that observer presence increased the frequency of pupil-teacher interchanges but had no effect on the ratio of appropriate to inappropriate behavior for either teacher or children. The maximum length of each phase was 10 days, yet there was not the slightest suggestion of habituation to the observer.

The final classroom study of observer presence effects was described earlier (Kent, Fisher, & O'Leary, 1974). To reiterate, Kent found a significant interaction between experimental phase (baseline vs. intervention) and intrusiveness of observation using inappropriate child behavior as the dependent variable. The covert observation was done via one-way mirror and the counter-balanced schedule of observations lasted 29 days (11 baseline,

18 treatment). No evidence of habituation was found. Two replications of this study are presently being conducted (Kent, in preparation; Weinrott, Walker, & Hops, in preparation).

A number of investigators have treated reactivity as negligible (Heyns & Lippett, 1954; Kerlinger, 1964) or as rapidly dissipating (Barker & Wright, 1955; Medly & Mitzel, 1963; Selltiz, Jahoda, Deutsch, & Cook, 1959; Werry & Quay, 1969; Wright, 1967). It appears that neither of these positions is founded upon solid, empirical evidence. It is clear that observer effects do exist and may persist over a period of weeks or longer. But, it is not at all apparent how they vary with respect to setting, length of observation session, artificial constraints placed on subjects, or treatment status. Other relevant factors include conspicuousness of the observer, individual differences of subjects, personal attributes of the observer, and rationale for the observation.

The following precautions were taken to minimize the potential effects of observer presence in the present study: (a) all observers were women; in elementary school, both children and teachers are unaccustomed to adult male presence in the classroom; (b) observers did not interact with children at all and kept conversation with the teacher to the bare minimum; (c) observers remained stationary except when the target child was not visible; (d) the observer rotated through peers at alternating intervals, which decreased the likelihood that the target child would detect his status as such; (e) observers were instructed to avoid wearing bright clothes or heavy make-up; (f) teachers were given a reasonable rationale for the observations (i.e., quasi-independent evaluation of the training program was preferred by the granting agency; control subjects were told that the study was to examine

interaction patterns of acting-out and withdrawn children; (g) a teacher's grade in the course was not contingent upon her implementing any programs in her classroom, thereby reducing demand characteristics; (h) no changes in the class timetable or activity structure were required; (i) young children constituted the sample, as opposed to older students, who may well have shown greater reactivity; (j) two pre-baseline observations were held so that the potential novelty surrounding the observer's apparatus (e.g., audio-pacer) would not affect the baseline; (k) the observations began in January, after a routine had been established, and many outside observers (e.g., administrators) had already visited.

Method: Experiment II

Subjects and Recruiting of Sample

Ss were 50 female elementary school teachers of regular grade 1-3 classes, 40 of whom participated in Experiment I. These 40 were recruited for participation in a three-credit behavior modification training program at McGill University. Key administrative personnel in five Montreal area school commissions and 10 private schools were contacted and informed of the nature of the proposed research. Each agreed to distribute a course description to all eligible teachers under their jurisdiction. Approximately 2,000 announcements were distributed, producing 204 respondents who identified themselves by returning a coupon requesting further information. Of this number, 61 were ineligible because they did not meet the criterion of being a regular 1-3 teacher in an English-speaking class. The remaining 143 respondents were mailed a letter requesting that they complete a series of behavior checklists pertaining to three children in the class--the most withdrawn, the most disruptive, and the most distractible. It was communicated that this was necessary for planning the course curriculum in a manner that would be tailor-made to the immediate needs of the participants. Also, it would help the instructor to place the participants in sections where a heterogeneity of problems would be presented. Completion of the checklist was not tantamount to an offer of admission into the program. No association was made between written responses and eligibility for entry. This precaution was intended to reduce response bias on the part of teachers who might have made their children "look bad" in order to increase the likelihood of acceptance.

Upon receipt of the behavior checklists a teacher was either rejected

because no child was rated as sufficiently deviant or she proceeded to the next phase of the selection process. The perceived severity of the child's handicap had to be rated as "a major problem requiring professional intervention" or "a minor problem worthy of a casual, short-term treatment program." In addition, a socially withdrawn child had to display two of the following characteristics at criterion activity levels:

volunteers in class	rarely or never
daydreams	quite often or always
initiates conversation with peers	rarely or never

A distractible child had to receive three of the following ratings:

finishes things he starts	rarely or never
out of seat	quite often or always
disturbs others	slightly more than average, quite often, or always
restless, fighting	quite often

A disruptive child needed two of the following ratings to qualify:

teases or interferes with other children	quite often or always
seeks attention of teacher	quite often or always
strikes back with aggressive behavior when teased or interfered with	quite often or always

Finally, a child must have fulfilled one of the following conditions:

1. Been the subject of a non-compulsory, teacher-initiated parent conference;
2. Been discussed by teacher with principal at least once;
3. Been discussed by teacher with counselor at least once;
4. Been referred for psychological testing by school.

All teachers who had at least one child who met minimum standards were telephoned and informed that a one-hour observation by E would occur in order to ascertain whether her target pupil(s) would be likely to benefit from behavioral intervention. Teachers were also notified of three additional contingencies which would be applied in the event they were accepted. These

included the purchase of a text book (Becker, Engelman, & Thomas, 1971); agreement to 20 in-class observations and attendance at all group sessions. Of 84 candidates who were telephoned, all but one agreed to the preliminary observation and supplementary contingencies. E conducted all pre-baseline observations using a classroom coding system to be described in subsequent sections. While E focused on only one or two eligible children and randomly selected peers, he asked the teacher to identify all three pupils she had targeted on the checklists and any others she may have been concerned with. It was hypothesized that if the teacher believed E was tracking the behavior of at least three children, she would be less inclined to influence the behavior of any single child in a manner that would improve her prospects of admission. Ratios between target child behavior and composite peer responses were computed using percentage of intervals in which a behavior occurred as the dependent variable. To be accepted, a "socially withdrawn" child was required to fulfill three of the following four conditions:

appropriate interaction with peers	maximum 1/2 peer norm
volunteering	maximum 1/2 peer norm
initiation to teacher	maximum 1/2 peer norm
looking around or self-stimulation	minimum 2 times peer norm

Distractible and disruptive children seemed almost indistinct from one another according to the teacher reports and, to a somewhat lesser extent, the preliminary observation. To be eligible, a target child had to have a total deviance score which was at least twice that of the composite peer norm. The total deviance score was derived by taking the total number of intervals in which one to 10 negative behaviors occurred and dividing it by the total number of intervals. In borderline cases the subject was either observed a second time or was rejected immediately. Following the preliminary screening, teachers were told they would be notified of their admission status within a few days.

Eligible teachers were admitted consecutively until quotas of 20 socially withdrawn and 20 disruptive-distractible children were filled. Two teachers in each category were assigned to a waiting list in the event of drop-outs.

Forty teachers were offered admission into the course and were asked to complete the necessary registration forms and provide travel directions to their schools, along with a timetable showing all periods when the target child was in the classroom. All invited applicants responded favorably to the offer and registered within the allotted time period.

Five teachers per pupil type were assigned to each of the four conditions described in Experiment I. Assignment was random following certain provisions for those who could attend group meetings only on particular days and for those schools which were represented by more than one teacher. In this instance, all participants from one school were assigned to the same experimental condition. There were no differences between groups on the number of teacher pairs (or in one case triads).

Ten matched control teachers were recruited by contacting appropriate administrative personnel in eight schools unrepresented in the experimental sample. The three behavior checklists were distributed, accompanied by a letter requesting cooperation with the research. Teachers were asked to complete the checklist, permit a maximum of two one-hour screening observations by E and, if their target child was deemed acceptable, fill out various questionnaires and consent to 20 additional in-class observations over the following five months. As compensation, each would receive two books at the completion of the study. Thirty-six teachers agreed to cooperate. Because of the ease in obtaining control subjects, it was possible to introduce an additional prerequisite which would control for possible differences in motivation

between experimental and control Ss. That is, all members of the latter group were to have satisfactorily completed at least one continuing education course within the last year or be currently registered. Due to an oversight on the part of E, this condition was fulfilled by only eight of the 10 control teachers. Any control subject who had received formal training in operant treatment of childhood disorders was rejected.

The total sample of 50 teachers was drawn from 34 different schools. Post hoc analyses revealed no differences between groups on number of years of teaching experience, size of class, amount of time contributed by third parties (i.e., student teachers, parent aides, etc.), and grade level.

Teachers in the four experimental conditions received 13 sessions of training in groups of 10 from January through May of 1974. Sessions were two hours long and were held on an irregular basis as noted below. Attendance was mandatory and was the only basis upon which evaluation was made. Three university credits were given on a pass-fail basis to all participants whose attendance was 100% for 13 sessions. Those unable to attend for reasons of ill health were required to arrange an individual meeting with the instructor in order to make up missed work.

Training Curriculum and Experimental Phases

Baseline (6 weeks; Phases I, II, and III of observational schedule) Session 1 (week 1). Completion of various dependent measures; presentation on defining behavior objectively, behavioral vs. medical models of psychopathology; pinpointing selected behaviors in the target child. Session 2 (week 2). Observation training including rationale for systematic naturalistic observation, instruction and practice in various data collection techniques. Groups E_3

and E_4 were assigned the task of daily data collection, the latter with an external monitor. Group E_1 received a "placebo" lecture on psychopharmacological intervention with hyperactive children. Session 3 (week 5). Alternative models of aggression, hyperactivity, and social withdrawal with particular emphasis on psychodynamic and ethological approaches.

While some would argue that inputs which occurred during baseline could be construed as a form of intervention (e.g., pinpointing, tracking), the principal hypothesis being tested required differential training and data collection during all phases of the experiment. A number of pre-baseline observations would have been advantageous but would also have imposed excessive demands on time and budget. Another consideration is that teachers were expected to attend meetings and to collect data for a period of 5-6 weeks before receiving treatment. It would have been somewhat difficult to hold them in abeyance for a more extended interval. The purpose of the somewhat unrelated presentation on alternative models (session 3) was to gather additional data and to remain in personal contact with the participants. Had this session not been included, a period of one month would have elapsed between meetings; such a large delay could have had an adverse effect on the quality of data collection and on expectations of improvement.

Intervention (10 Weeks)

Session 1 (week 7). Writing of behavioral objectives, introduction to applied behavior analysis; the rationale for and use of contingent attention; distribution of text and reading assignment.

Session 2 (week 8). Reinforcement, including types, how to, schedules of; problem analysis exercise which required identification of antecedent and

consequent stimuli, and selection of reinforcers; viewed film "Child Behavior Equals You" (National Film Board of Canada); reading assignment plus reinforcement survey for target child (Tharp & Wetzel, 1969).

Session 3 (week 9). Elimination of inappropriate and unacceptable behavior; time-out; ignoring; punishment, including why, why not, and how; viewed video tapes of time-out and ignoring; further work on problem analysis exercise with inclusion of extinction or punishment techniques; assignment: reading plus use of contingent attention.

Session 4 (week 10). Designing of individual programs for target children; presentation of modeling, symbolic modeling, and role-playing techniques for withdrawn children; point systems for distractible-disruptive children; assignment: reading, implement program, withdrawn children view symbolic modeling video tapes of social interaction and volunteering.

Session 5 (week 11). Review programs; presentation of the Good Behavior Game (Barrish, Saunders, & Wolf, 1969; Medland & Stachnik, 1972); use of group contingencies; addition of individual contingencies for withdrawn subjects; use of and distribution of work box (Patterson, Cobb, & Ray, 1972) for severely acting-out pupils; shaping, especially of academic skills, cueing; assignment: reading and continue program.

Session 6 (week 12). Double interlocking cross-over contingency for withdrawn children (Walker & Hops, 1973) modified for teacher administration; contingency contracting; using parents to deliver reinforcers; assignment: reading, continue program.

Session 7 (week 14). Modification of cross-over contingencies where necessary; elimination of self-stimulatory behavior; reduction of fear responses; assignment: reading, continue programs.

Session 8 (week 16). Setting up programs for other children in the same class (this had been encouraged all along, particularly in cases where the teacher did not consider the target child as her greatest problem); fading of reinforcers; shaping original responses.

Follow-up

During this phase teachers were advised to fade use of contrived reinforcers if they had not already done so and to gradually eliminate contingencies which required special tracking (i.e., those which would not likely be followed by another teacher). Inputs during follow-up were related to previously instituted behavior modification procedures; however, no new programs were introduced and only minor adjustments were made in several cases where reported progress was minimal. During follow-up, members of groups E_3 and E_4 were still required to collect daily data, the latter group continuing to utilize external monitors. Longer term follow-up was precluded by the close of school for the summer.

Session 1 (week 18). Tutoring programs; review fading of reinforcers; dealing with parents from a behavioral perspective; description and assignment of final reports on all programs implemented during the course. These were to be sumitted on ditto masters so each teacher in the entire study could receive a copy of every report.

Session 2 (week 20). Resources for further training and materials in behavior modification; description of an administrative strategy for implementing similar in-service training within a school commission; completion of post-intervention dependent measures.

Dependent Measures

Naturalistic observation data. The observation system employed herein was a modified version of the classroom coding procedures developed by Patterson, Cobb, and Ray (1972). The system has provisions for recording both target child and peer behavior, environmental responses to that behavior, and the type of ongoing activity prescribed by the teacher. Omnibus categories or composites were formed for Total Deviant Behavior, Disruptiveness, Distractibility, and Social Withdrawal. Each is a combination of two or more of the following 18 discrete responses:

Appropriate Behaviors

approval
compliance
appropriate interaction with teacher
appropriate interaction with peer
volunteer
initiation to or by teacher
laugh
attend

Inappropriate Behaviors

physical aggression
disapproval
high rate
noncompliance
inappropriate interaction with teacher
inappropriate interaction with peer
inappropriate locale
self-stimulation
look around
not attend

The 18 behaviors were dichotomously classified as either appropriate or inappropriate depending on their presumed acceptability in the classroom. Not only do these assignments have face validity, but a number of discrete responses lying within the Appropriate classification correlate positively with academic achievement for elementary school children, while several deemed

Inappropriate show negative correlations (Cobb, 1970, 1972). Rules for forming other omnibus categories (e.g., Disruptiveness) are presented with the results.

Appendix A consists of the coding manual used for observer training. It contains definitions for each of the 18 behaviors as well as complete instructions for conducting an observation. Briefly, the observer records the behavior of a target child and a systematically selected (on basis of seating arrangement) same-sex peer on alternating six-second intervals. While the target child remains the same, the observer rotates through all the peers, thereby providing normative data in the form of "composite peer" scores. Either the target child or a peer is therefore designated as the subject for an interval. Any responses to the behavior of this subject are also coded. Such responses may be emitted by the teacher or a peer. Again, the reader is referred to Appendix A for a more detailed description of the observation procedures and all other dependent measures.

Observers were required to provide information about the nature of the activity in which the class was engaged. Five conditions were defined as follows: (a) Structured: the teacher has provided clear guidelines for the children to follow in carrying out tasks. (b) Unstructured: the guidelines for the child's behavior are vague or unclear to the observer, i.e., the students can determine what they want to do in terms of academic activity. (c) Group: the class is involved as one unit in academic activity, e.g., the teacher lecturing, student reciting while others listen. Also, group applies to activities where the class is divided into several small units such as in reading or special projects. (d) Individual: the majority of the students are doing work by themselves at desks, e.g., art projects are being done by

each child. "Individual" applies even though the student asks for and receives help from other peers and/or teachers. (e) Transitional: the class is between activities, e.g., waiting for recess, lining up for lunch, returning from recess, teacher has indicated reading period but has provided no directions for the next activity.

The observation data served as dependent measures for testing hypotheses 2-8. Testing of Hypothesis 1 used the observation data as an independent or criterion variable against which the accuracy of teacher perception was appraised.

Walker Problem Behavior Identification Checklist (Walker, 1970). This is a 50-item checklist composed of operational statements about observable classroom behavior, which all teachers completed for target children during baseline and follow-up. Factor scores were obtained for acting out, withdrawal, distractibility, disturbed peer relations, and immaturity.

Summary reports. These are weekly ratings on a seven-point scale. Teachers whose target child was withdrawn completed a "withdrawal scale," while those whose target child was acting out provided a separate rating on dimensions of disruptiveness and distractibility. Ratings were obtained from teachers in group C_1 once during baseline and again at the termination of the study.

Behavior Vignettes Test (Heifitz, 1972). This is a 20-item multiple choice test that assesses knowledge of behavior modification principles and techniques of classroom management. This instrument had been successfully pilot-tested (Baker, Heifitz, & Pasick, 1973). It was administered to all teachers during baseline and following training.

Number of programs implemented. A breakdown of all behavior modification

programs implemented during the course of training was obtained from each teacher in groups E_1-E_4. A program was defined as: (a) pinpointing the problem, (b) setting a contingency, and (c) delivering the reinforcing or punishing consequence. The use of time-out was not included in this measure unless a reinforcing consequence was available for specified prosocial behaviors. Programs were categorized as either group or individual in nature with contingencies placed on either academic performance or classroom behavior.

Teacher global ratings. A post-treatment general rating of target child improvement measured on a four-point scale ("a great deal," "somewhat," "a little," or "not at all").

Expectation of improvement (Walter & Gilmore, 1973). An instrument requiring teachers or parents to estimate the probability of target child improvement. This estimate was obtained from experimental teachers at the beginning of each group session.

Cost analysis. This measure consists of cost estimates for implementation of the present teacher training program excluding research components. Such costs were measured both monetarily and in terms of time expenditures for a psychologist and participating teachers. The difficulty in assigning a dollar value to the benefits accrued precludes a true cost-benefit analysis. The cost data supplied here are useful only insofar as the present intervention can be compared to alternative models with similar objectives.

Results: Experiment II

Teacher Attendance

Each of the 40 teachers was expected to attend 13 group sessions. The total number of possible attendances was 520. Only 21 absences were recorded, thus producing an overall attendance of 96%. Of the 21 absences, six were accounted for by one teacher (#26) who ultimately did not receive credit for participating in the program. However, all data were collected for this subject and these were included in the analysis of group E_2.[9] No other subject was absent more than twice. None of the original 50 teachers in the study dropped out.

Dropouts of Target Children

One disruptive/distractible target child (#45) was transferred to a special education class just prior to the onset of intervention. Baseline data were not retained. Therefore, data for the acting-out subsample from group E_4 were obtained from four subjects only. It is perhaps worth noting that teacher #45 continued to participate fully in the absence of a specific target child.

Subject #16, an acting-out child in group E_1 moved out of the area after intervention had been completed. His data are included in all analyses.

---

[9] It was decided a priori that if the score on the second administration of the Behavior Vignettes Test was lower than one half a standard deviation below the mean for trained teachers then data for Subject #26 were to be excluded. This teacher had a score of 7 on the pre-test and 12 on the post-test. Since the post-treatment mean for groups E_1-E_4 was 11.72, she was considered to have learned basic social learning principles and techniques despite a poor record of attendance.

Follow-up observation data for the acting-out subsample of group E_1 were obtained from four subjects only.

There were no dropouts among socially withdrawn children.

Observer Agreement

A total of 148 reliability checks were made during the course of data collection. These occurred at a rate of two per week per observer except during the final week when checks were eliminated.

The overall mean reliability was .91.

Percentages of observer agreement for each code category are presented in Table 2.1.

Table 2.1

Percentages of Observer Agreement for Each Code Category

Code		Percent agreement	Code		Percent agreement
AP	Approval	92	PA	Physical aggression	98
CO	Comply	89	DI	Disapproval	83
T+	Appropriate interaction with teacher	93	HR	High rate	81
			NC	Noncomply	84
P+	Appropriate interaction with peer	89	T-	Inappropriate interaction with teacher	87
VO	Volunteering	83	P-	Inappropriate interaction with peer	87
IT	Initiation to/by teacher	88	IL	Inappropriate locale	92
LA	Laugh	84	SS	Self-stimulation	81
AT	Attend	95	LO	Looking around	83
			NA	Not attend	86

Because interval-by-interval reliabilities were consistently high, it was deemed unnecessary to compute correlation coefficients for each dependent variable (composite score).

Transformation of Observation Data.

Observation data (frequencies) were converted to proportion scores for each of the 18 behavior categories. This was done to compensate for slight variation in the length of observation sessions. Arcsin transformations were performed in order to stabalize the variances. Following this, a category mean and standard deviation were computed for the pooled composite peer data from Phases I-III.[10] Using the formula for standard score conversion,

$$z = \frac{X-\bar{x}}{\sigma}$$

the mean and standard deviation for the peer sample were substituted, along with the transformed portion "x" for the target subject or group. The obtained z scores represent relative or comparative performance. Table 2.2 presents the corresponding proportions for z score values. A thorough explanation of the rationale and procedure for standard score conversion of observation data is presented in Appendix B.

Selection of Dependent Variables

Three dependent variables were relevant to the testing of Hypothesis 1. These include "disruptiveness," "distractibility," and "social withdrawal."

---

[10] Peer data for Phases IV-VI were not included in calculation of norms because peer behavior under these conditions could well have been affected by any of the following contingencies (Walker & Hops, 1975): (1) a target subject shares individually earned rewards with his peers; (2) peers are encouraged to make social responses to a target child in order to assist him in meeting the criterion for group rewards; (3) one child is the subject of an intervention but is not identified as such; here the treatment procedures and rewards involve all members of the class; (4) both the target child and peers are participating in treatment and are required to cooperate in order to fulfill reinforcement criteria.

Table 2.2

Proportions of Behavior Resulting in a Given z Score

z	Appropriate								Inappropriate									
	AP	CO	T+	P+	VO	IT	LA	AT	PA	DI	HR	NC	T-	P-	IL	SS	LO	NA
3.75	.01	.15	.18	.46	.30	.22	.05	.97	.03	.02	.12	.03	.09	.23	.07	.16	.32	.16
3.50	.01	.14	.16	.43	.28	.20	.04	.96	.03	.02	.11	.03	.08	.21	.07	.15	.30	.15
3.25	.01	.13	.15	.40	.25	.18	.04	.94	.03	.02	.10	.03	.07	.19	.06	.13	.28	.13
3.00	.01	.11	.14	.36	.23	.17	.03	.93	.02	.02	.09	.02	.06	.18	.05	.12	.26	.12
2.75	.01	.10	.12	.33	.20	.15	.03	.91	.02	.01	.08	.02	.05	.16	.04	.11	.24	.11
2.50	.00	.09	.11	.30	.18	.14	.02	.89	.02	.01	.07	.02	.05	.14	.04	.09	.22	.09
2.25		.08	.10	.26	.16	.12	.02	.87	.01	.01	.06	.01	.04	.12	.03	.08	.20	.08
2.00		.07	.09	.24	.14	.11	.02	.85	.01	.01	.05	.01	.03	.11	.03	.07	.18	.07
1.75		.07	.08	.21	.12	.09	.01	.83	.01	.01	.04	.01	.03	.09	.02	.06	.16	.06
1.50		.06	.07	.18	.10	.08	.01	.80	.01	.01	.04	.01	.02	.08	.02	.05	.14	.05
1.25		.05	.06	.16	.08	.07	.01	.78	.01	.00	.03	.01	.02	.07	.01	.04	.13	.04
1.00		.04	.05	.13	.07	.06	.01	.75	.00		.02	.01	.01	.06	.01	.04	.11	.03
.75		.03	.04	.11	.05	.05	.01	.72			.02	.00	.01	.05	.01	.03	.10	.03
.50		.03	.03	.09	.04	.04	.00	.69			.01		.01	.04	.01	.02	.08	.02
.25		.02	.03	.07	.03	.03		.66			.01		.00	.03	.00	.02	.07	.02
.00		.02	.02	.06	.02	.02		.64			.01			.02		.01	.06	.01
- .25		.01	.02	.04	.01	.02		.60			.00			.01		.01	.05	.01
- .50		.01	.01	.03	.01	.01		.57						.01		.00	.04	.00
- .75		.01	.01	.02	.00	.01		.54						.01			.03	
-1.00		.00	.00	.01		.00		.51						.00			.02	
-1.25				.01				.48									.02	
-1.50				.00				.45									.01	
-1.75								.42									.01	
-2.00								.39									.00	
-2.25								.36										
-2.50								.33										
-2.75								.30										

Teachers whose target child was defined as "acting out" submitted weekly ratings on the first two variables, while those whose target child was "socially withdrawn" followed the same procedure on the third variable. The reader is reminded that ratings were made on a seven-point scale with "7" designated as the most extreme condition. The selection of composites, or parallel constructs based on direct observation, was carried out in a somewhat different manner for each of the three variables to be discussed below and several others to be introduced in later sections.

Disruptiveness: $(\underline{PA}) + (\underline{DI}) + (\underline{T-}) + (\underline{P-})$

This variable was defined empirically as that combination of deviant behaviors to which the environment (i.e., teacher or peers) responded with an overall probability greater than .5 during baseline. A ratio was computed by dividing the total number of intervals in which a particular behavior occurred into the number of intervals in which any response (i.e., positive, negative, or neutral) to that behavior was recorded. For example, if "inappropriate interaction with teacher" (e.g., calling out) occurred 200 times for "i" children over "j" observations and teacher reprimands, peer laughter or any other reaction followed, 140 of these incidents within six seconds, then the disruptive ratio would simply be $\frac{140}{200}$ or .70.

Multiplying this ratio by 100 yields a percentage of teacher or peer reaction to each class of behavior. These are presented in Table 2.3 for all deviant categories. A further breakdown comparing target children to peers indicates that the environment is remarkably stable in its response probability to a particular discrete behavior. Given a particular behavior, the probability of it disrupting others is about the same regardless of who emitted it.

Generally, the disruptive target children were responded to with only a slightly lower probability than their peers. Any category for which the combined target/peer disruptiveness percentage exceeded 50 was included in the composite. Behaviors fulfilling this requirement were physical aggression (60%), disapproval (58%), inappropriate interaction with teacher (51%), and inappropriate interaction with peer (82%). The next most disruptive behavior, "inappropriate locale," showed far less success in eliciting environmental responses (18%). Since the term "disruptive" pertains to interference with others, it would appear that the present manner of defining it yields a variable of good construct validity.

Table 2.3

Environmental Responses to Deviant Behaviors

	% environmental response		
	To target children	To peers	Combined
PA	58.58	62.86	59.88*
DI	55.47	62.50	57.89*
HR	9.58	9.80	9.65
NC	12.26	13.13	12.47
T-	49.38	56.45	50.55*
P-	79.48	85.66	82.11*
IL	17.69	18.26	17.84
SS	1.32	0.56	1.08
LQ	2.14	0.74	1.48
NA	4.57	5.08	4.70

* $p < .05$

Distractibility: (HR) + (P-) + (IL) + (SS) + (LO) + (NA)

This variable includes those discrete responses designated as "distractible" in Experiment I. The combination was formulated arbitrarily, but consensus as to its content validity was obtained verbally by the 40 experimental teachers. Component categories include high rate, inappropriate interaction with peer, inappropriate locale, self-stimulation, looking around, and not attending. The reader should refer to the coding manual in Appendix A for complete definitions and relevant examples.

Social Withdrawal: (P+) + (VO) + (IT) - (SS) - (LO)

The combination for withdrawal represents those behaviors upon which sample selection was based: appropriate interaction with peer, volunteers, initiation to teacher, self-stimulation, and looking around. Unlike the previous two composites, withdrawal consists of both appropriate and deviant behavior codes, thus accounting for the minus signs. Target children were deficient on the first three categories and above the normal level for the last two. t-tests for correlated samples showed that each of the five responses discriminated target subjects from peers during baseline Phases I and III. Table 2.4 presents the results of these analyses.

Testing for a Tracking Effect

In order to examine the relationship between teacher perceptions (i.e., ratings) of behavior change and independently observed behavior levels, the weekly ratings on the seven-point scale were correlated with observation data. Eighteen pairs of summary ratings and observation composite scores were formed

for each teacher-student dyad by correlating the rating obtained at the beginning of a week with the composite scores from the corresponding (i.e., most recent) observation.[11] Observations seldom preceded a summary report by more than four school days.

Table 2.4

Behaviors Discriminating Withdrawn Children from Normal Peers

Code category	df	t (one-tailed
P+	24	5.642**
VO	24	4.085**
IT	24	4.525**
SS	24	3.514**
LO	24	5.946**

** $p < .01$

While there is little reason to expect any relationship to exist between one hour of classroom observation and a weekly summary report, it was assumed that reported improvement would yield moderately strong correlations over the course of 18 observations in cases where behavior change was, in fact, demonstrated. Similarly, reports indicating little or no progress would be expected to correlate significantly with observed behavior levels which remain relatively stable over the course of treatment.

Ratings of "disruptiveness" were correlated with the composite score (PA) + (DI) + (T-) + (P-); "distractibility" with (HR) + (P-) + (IL) + (SS) + (LO) + (NA), and "social withdrawal" with (P+) + (VO) + (IT) - (SS) - (LO). Following

[11] Data from the demand baseline condition was excluded from this analysis.

computation of Pearson product-moment correlation coefficients, a standard score transformation was performed to normalize the distribution of "r." This rendered the sampling distribution of z scores independent of the magnitude of the correlation coefficient "r." Table 2.5 presents correlations and z scores on each dependent variable for each subject. Notice that correlations for withdrawal are mostly negative. This is because the combination of behaviors yields higher scores as the child improves. For acting-out subjects, improvement on either the disruptiveness or distractibility composite is denoted by a decrease in the score. A separate one-way ANOVA for each of the three sets of z scores was performed to test for group differences in convergent validity.

Group means and standard deviations for disruptiveness are presented in Table 2.6. One can easily detect that no differences in mean levels existed between E_1, E_2, and E_3. Despite a higher z score of .457 for E_4, no main effect for groups was obtained ($\underline{F}$ = 0.433, $\underline{df}$ = 3, 15; n.s.). (See Table 2.7.) While it appears that the group receiving the maximum amount of observation training and monitoring was somewhat more accurate in rating disruptiveness, this difference did not approach significance.

Turning to "distractibility," one can see from Table 2.8 that the order of group means corresponds very nearly to the amount of observational input, with only E_2 and E_3 showing no discernable differences from one another. Although the lowest score of .282 for E_1 appears substantially lower than .565 for E_4, this difference was not statistically significant ($\underline{F}$ = .657, $\underline{df}$ = 3, 13; n.s.). A very high standard deviation of .482 for E_4 helps account for the absence of a main effect (see Table 2.9). Also, the small number of subjects in each group mitigates the attaining of statistical significance.

Table 2.5

z Transformed r's

Disruptiveness			Distractibility			Withdrawal		
Group/ S no.	r	z	Group/ S no.	r	z	Group/ S no.	r	z
1/5	0.2076	0.2106	1/5	0.1883	0.1905	1/0	0.4398*	-0.4719
1/6	-0.0356	-0.0356	1/6	-0.1096	-0.1100	1/1	-0.7573**	-0.9899
1/7	0.4575*	0.4942	1/7	0.4427*	0.4755	1/2	-0.5486**	-0.6163
1/8	0.0670	0.0671	1/8	0.5044*	0.5551	1/3	-0.4501*	-0.4848
1/9	0.3630	0.3804	1/9	0.2893	0.2978	1/4	-0.4798*	-0.5228
2/5	0.4280*	0.4574	2/5	0.3766	0.3961	2/0	0.0383	0.0383
2/6	0.2299	0.2341	2/6	0.3354	0.3489	2/1	-0.3584	-0.3750
2/7	0.3921	0.4143	2/7	0.6903**	0.8486	2/2	-0.2010	-0.2037
2/8	0.2216	0.2254	2/8	0.3875	0.4088	2/3	0.1048	0.1052
2/9	-0.1275	-0.1282	2/9	0.0583	0.0584	2/4	-0.4275*	-0.4569
3/5	0.6521**	0.7789	3/5	0.4522*	0.4874	3/0	-0.4041*	-0.4286
3/6	0.0806	0.0808	3/6	0.4553*	0.4914	3/1	-0.1541	-0.1554
3/7	0.5232*	0.5807	3/7	0.4733*	0.5144	3/2	-0.1589	-0.1602
3/8	0.1885	0.1908	3/8	0.3592	0.3760	3/3	0.3649	0.3826
3/9	-0.3003	-0.3099	3/9	0.1754	0.1773	3/4	-0.2182	-0.2218
4/6	0.3712	0.3898	4/6	0.7644**	1.0068	4/0	-0.2385	-0.2432
4/7	0.7307**	0.9303	4/7	0.6438**	0.7646	4/1	-0.6745**	-0.8189
4/8	-0.1175	-0.1180	4/8	-0.1141	-0.1146	4/2	-0.3293	-0.3420
4/9	0.5557**	0.6266	4/9	0.5403**	0.6046	4/3	-0.3987	-0.4221
						4/4	-0.8704**	-1.3346

* $p < .05$
** $p < .01$

Table 2.6

Mean z Scores and Standard Deviations for Disruptiveness

	E_1	E_2	E_3	E_4
Mean	0.22334	0.24060	0.26426	0.45717
Standard deviation	0.21775	0.23101	0.42828	0.44269

Table 2.7

Analysis of Variance for Disruptiveness

Source	SS	DF	MS	F
Mean	1.65307	1	1.65307	14.37675
Group (A)	0.14944	3	0.04981	0.43321
Error	1.72473	15	0.11498	

Table 2.8

Mean z Scores and Standard Deviations for Distractibility

	E_1	E_2	E_3	E_4
Mean	0.28178	0.41216	0.40930	0.56535
Standard deviation	0.26187	0.28280	0.14037	0.48251

Table 2.9

Analysis of Variance for Distractibility

Source	SS	DF	MS	F
Mean	3.27551	1	3.27551	35.82466
Group (A)	0.17875	3	0.05958	0.65168
Error	1.37148	15	0.09143	

Across both dependent variables for "acting-out" children, it appears that observation training may account for slight increases in the accuracy of teacher summary reports since group E_1, which received no such training, showed the lowest transformed correlation on both variables. It also seems that for a few teachers, the introduction of an external monitor may improve the accuracy of perception. Group E_4 did show the highest degree of correspondence between ratings and behavior but also displayed markedly higher variability. The most parsimonious conclusion to be drawn from this analysis is that none of the three components of observational input (i.e., training, daily data collection, external monitoring) was capable of improving the reliability of teacher ratings. Still, F ratios were greater than 1.0 and over half the correlation coefficients for E_3 and E_4 were significant (see Table 2.5). Only five significant correlations were found of the 20 possible for E_1 and E_2.

An examination of the composite for social withdrawal vs. ratings of withdrawn subjects indicates that the more conservative interpretation may be the most appropriate. Group means in Table 2.10 show that groups E_1 (-.617) and E_4 (-.632) were clearly higher than those for E_2 (-.178) and E_3 (-.117). The analysis of variance (see Table 2.11) yielded a significant effect for groups

Table 2.10

Mean z Scores and Standard Deviations for Withdrawal

	E_1	E_2	E_3	E_4
Mean	-0.61714	-0.17842	-0.11668	-0.38610
Standard deviation	0.21590	0.24710	0,30042	0.44942

Table 2.11

Analysis of Variance for Withdrawal

Source	SS	DF	MS	F
Mean	2.98145	1	2.98145	29.82166
Group (A)	1.14821	3	0.38274	3.82827*
Error	1.59962	16	0.09998	

* $p < .05$

($F = 3.828$, $df = 3, 16$; p .05). Further evidence discounting a tracking effect comes from inspection of the standard deviation. Not only did the group receiving no observation training whatever show a relatively high mean, but this was accompanied by little variability. Such was not the case for E_4. Group E_1 also showed significant correlations for each of its five members, while no other group produced more than two (see Table 2.5). Clearly, for socially withdrawn children, there appears to be no tracking effect analogous to that exhibited in the laboratory studies.

A naturalistic tracking effect could be manifested in a manner other than the accuracy of perception. Three additional sources of inter-group variation were investigated: teacher impressions of the target child, the actual behavior

of the child, and various dimensions of teacher attention. The analogue studies permitted only the summary reports to vary as a function of observation training. Child behavior in the laboratory was predetermined and non-participant judges could not actively interact with the stimulus children. Yet, the indirect influence of observation training on the classroom behavior of teacher or students was a distinct possibility.

Neither the summary reports nor any measure of target child behavior produced strong evidence of a tracking effect. However, there appears to have been some impact of daily data collection on teacher attention to withdrawn children. The following appraisal of the intervention includes the detailed results.

Evaluation of the Intervention

One of the unique properties of the present experimental design is a provision for making three types of comparisons: (a) intra-subject or intra-group, (b) inter-group (e.g., treatment vs. control), and (c) normative (e.g., target children vs. peers). Where appropriate, status of dependent variables will be evaluated using all three standards.

Observation Data

To test for serial dependency among data points, autocorrelations (lags 1-10) were computed for each of four dependent variables (total deviant behavior excluding "looking around," disruptiveness, distractibility, and social withdrawal). Results for both group and individual data showed generally non-significant lag_1, autocorrelation coefficients, thus indicating little serial dependence from one observation to the next. This finding, coupled with the relatively small number of observations per phase, precluded the use of an

interrupted time-series procedure (Box & Tiao, 1965; Glass, Willson, & Gottman, 1973; Jones, Vaught, & Reid, 1973).

A generalizability model (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) was also considered. This would have involved a components of variance analysis of four facets (target/peer, group, phase, and activity). However, the large number of zero and near-zero entries due to a particular activity not occurring within a phase would have yielded artificially inflated z scores.

Consequently, each of the four dependent variables was subjected to a 2 (target/peer) x 5 (group) x 6 (phase) analysis of variance for repeated measures. Intervention was divided into early treatment (Phase IV) and later treatment (Phase V) for some analyses. Results for three variables relevant to acting-out children will be presented first.

Total deviant behavior (excluding "looking around"). t-tests for correlated samples revealed that all 10 categories of deviant behavior except "looking around" discriminated acting-out target children from their peers during baseline Phases I and III (see Table 2.12). Means and standard deviations for the nine-category combination of summed z scores are presented in Table 2.13. The analysis of variance yielded significant main effects for all three factors (see Table 2.14). Orthogonal comparisons relevant to specific hypotheses were subsequently performed. Figure 2.1 shows z scores by observation for acting-out target children in experimental groups, peers, and matched controls.

The overall mean for the four experimental groups during Phases I and III was 16.908.[12] During the demand procedure the level increased slightly to

---

[12] For purposes of comparing the present sample with those reported elsewhere, the natural baseline mean and standard deviation for percentage of time off-task (excluding "looking around") were 28.24 and 16.54, respectively.

Table 2.12

Discriminant Validity of Deviant Behaviors

Response category		t
Physical aggression	(PA)	3.069**
Disapproval	(DI)	3.467**
High rate	(HR)	4.218**
Noncompliance	(NC)	5.477**
Inappropriate interaction with teacher	(T-)	3.147**
Inappropriate interaction with peer	(P-)	2.278*
Inappropriate locale	(IL)	2.188*
Self-stimulation	(SS)	4.409**
Looking around	(LO)	1.189
Not attend	(NA)	4.160**

* $p < .05$ (one-tailed, $df = 23$)
** $p < .01$

16.963, clearly a non-significant change ($F = 0.00$, $df = 1, 190$; n.s.). The instruction to "make the target child appear quiet and cooperative" was not at all effective in reducing deviant behavior. Yet, it is obvious that teachers did attempt to exert additional control. Teachers in groups E_1-E_4 increased their number of responses per observation to the target child by 36.70% from Baseline I to the demand condition, while control teachers, who did not receive any instructions, showed a small decrease (see Table 2.15). The return to natural baseline was accompanied by a 13.24% reduction in teacher attention for groups E_1-E_4. In order to demonstrate that the increase in contact with the target child was not accomplished by a concomitant rise in overall teacher-pupil interactions, a measure of relative attention was computed. Table 2.16 shows the ratio of teacher responses to target children divided by her attention to peers. First, it is obvious that target children garner disproportionate amounts of attention, approximately 50% more than their

Table 2.13

Mean z Scores for Total Deviant Behavior (Excluding Looking Around)

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	16.91679	12.69300	10.11440	15.71559	8.21320	3.72540
	Peer	7.25240	2.22840	4.20820	7.26340	0.09360	1.86400
E_2	Target	18.50079	14.95159	17.80278	11.3564	8.76400	6.38059
	Peer	3.06620	4.33740	7.69860	1.03540	4.98199	3.05660
E_3	Target	27.22319	27.02258	16.36238	18.52838	12.24960	10.04219
	Peer	14.18539	11.90959	10.50419	9.53799	6.53860	7.65559
E_4	Target	18.11998	12.24075	8.85674	9.70900	7.47050	1.93100
	Peer	1.52770	1.00320	1.10550	-0.07800	2.00550	-0.79650
C_1	Target	19.66658	17.55199	11.20039	17.57179	16.26259	18.71059
	Peer	8.99359	9.22779	2.33100	7.07939	7.71839	5.37199

Table 2.13 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	11.71287	4.83993	9.81108	8.65618	6.95084	7.02148
	Peer	5.49232	5.53761	5.45971	4.61172	5.22588	5.56338
E_2	Target	12.60442	11.66020	12.44393	7.91591	6.95917	12.94440
	Peer	2.16156	5.00939	4.50274	5.45811	5.19398	5.19509
E_3	Target	10.35508	10.01629	5.18987	12.73524	6.62952	6.42409
	Peer	8.77103	7.45729	3.34918	6.43094	8.60261	6.72986
E_4	Target	7.65246	6.37569	0.82197	6.68811	6.66566	7.45050
	Peer	4.00320	4.16240	5.21160	7.66410	4.37060	3.81770
C_1	Target	12.01743	9.49284	6.06109	8.35064	10.95585	11.36235
	Peer	12.30834	11.77744	5.59197	11.31044	8.43620	7.25007

Table 2.14

Analysis of Variance for Total Deviant Behavior
(Excluding Looking Around)

Source	SS	DF	MS	F
Mean	25673.51172	1	25673.51172	98.27162
Target child/peer (A)	5487.30469	1	5487.30469	21.00400***
Group (B)	2861.18359	4	715.29590	2.73797*
A x B	53.51953	4	13.37988	0.05121
Error	9927.51953	38	261.25049	
Phase (C)	1808.85547	5	316.77100	13.90740***
A x C	576.43750	5	115.28749	4.43194**
B x C	1285.85153	20	64.29257	2.47157**
A x B x C	402.88672	20	20.14433	0.77440
Error	4942.44141	190	26.01285	

* $p < .05$

** $p < .01$

*** $p < .001$

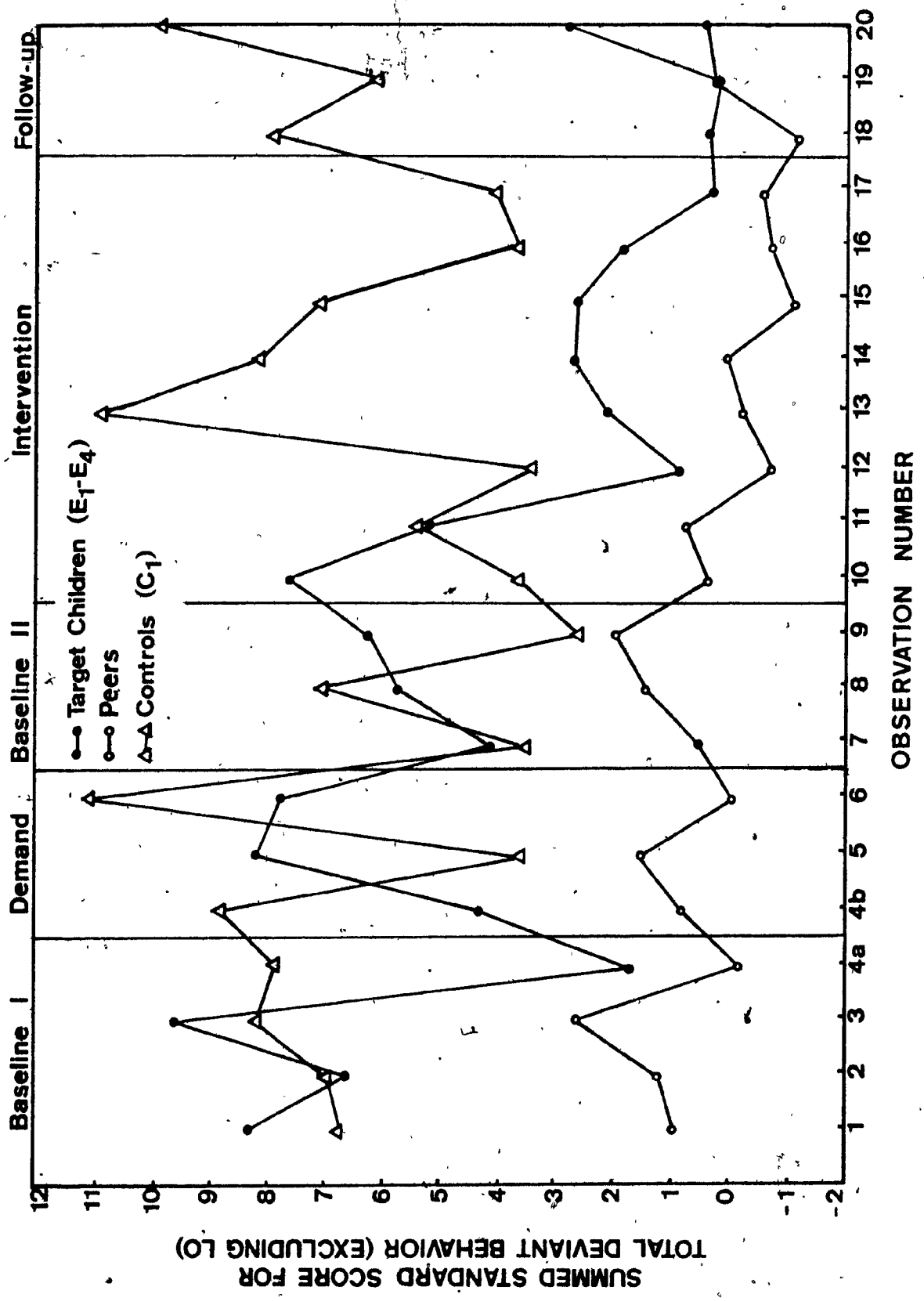


Figure 2.1. z Scores for Total Deviant Behavior (excluding LO) for Treated Target Children, Peers, and Matched Controls.

Table 2.15

Percentage of Intervals in Which Teacher Attended to Target Child

	Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phases IV-V Intervention	Phase VI Follow-up
E_1	6.83	10.51	9.37	8.64	7.65
E_2	8.60	13.99	11.65	12.08	12.79
E_3	10.23	9.26	10.48	9.95	10.38
E_4	6.64	12.27	8.86	8.82	8.67
C_1	14.23	9.29	9.39	7.91	8.69
Mean (E_1-E_4)	8.07	11.50	10.09	9.87	9.87

Table 2.16

Ratio of Teacher Attention Delivered to Acting-Out Target Children and Peers

	Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phases IV-V Intervention	Phase VI Follow-up
E_1	1.12	1.89	1.23	1.55	1.08
E_2	2.17	2.31	1.86	1.98	1.67
E_3	1.46	1.21	1.78	1.31	1.42
E_4	1.35	1.77	1.65	1.37	1.50
C_1	1.53	1.40	1.57	1.14	.94

peers. Notice also that three of the four experimental groups (excluding E_3) show increases in relative attention directed toward the acting-out child during the demand condition, while group C_1 shows a slight decrease. The level

attained by each of the three experimental groups which showed increases was higher than any other point in the study. This suggests that the change in teacher attention was due specifically to the instructional set. Further evidence confirming the impact of demand characteristics on teacher attention comes from the control group which neither participated in the demand condition nor exhibited much variability in relative attention between baseline phases I-III.

The natural baseline mean for experimental groups was 17.387. This was reduced to 9.259 in late intervention. This substantial decrease in deviant behavior is the equivalent of a one-standard deviation improvement in each of the nine deviant behaviors (F = 27.592, df = 1, 190; $p < .01$). Intervention was successful in significantly reducing noxious behavior in each of the four experimental groups. Visual inspection of Table 2.13 reveals no systematic differences which could be accounted for by observation training, daily data collection, or monitoring. Group C_1 showed no discernable improvement over time as indicated by a comparison of total baseline with Phase VI observations (F = .002, df = 1, 190; n.s.).

It is obvious from an examination of the follow-up data (Phase VI) that improvement continued during fading of contrived reinforcers and the more mechanistic aspects of particular programs. Follow-up data for E_1-E_4 produced an overall mean of 5.709 compared to 18.711 for C_1. This difference was highly significant (F = 17.099, df = 1, 190; $p < .001$).

The target child mean of 5.709 during follow-up remained higher than the corresponding peer score of 1.850. However, this difference did not attain statistical significance (F = 2.399, df = 1, 190; $p < .25$). Thus, while somewhat more deviant, the target children were behaving within the normal range.

Such was not the case for control subjects ($\bar{x}$ = 18.711), who still differed from their peers ($\bar{x}$ = 5.372) in Phase VI (F = 17.099, df = 1, 190; $p < .01$).

None of the aforementioned comparisons performed on this dependent variable indicated greater behavior change as a function of observation training, daily tracking, or monitoring. All groups (E_1-E_4) showed a somewhat similar pattern of improvement. Group E_4, which received the optimal amount of observational input, did not improve as rapidly as the other groups, although their follow-up mean of 1.931 was the lowest. On the other hand, E_1, which merely applied the treatment procedures without prior or subsequent tracking, appears to have performed similarly well.

There also appears to have been no effect of observation training on teacher reinforcement patterns, although several rather striking differences between experimental and control groups emerged. First, the percentage of intervals in which the teacher responded to the target children did not vary appreciably as a function of group status or intervention. Teachers in E_1-E_4 attended to target subjects in 9.00% of the intervals in Phases I and III and increased only to 9.87% during intervention, remaining stable thereafter. No differences among these groups were detected (see Table 2.15). The control group baseline mean (Phases I-III) was 11.25%, intervention 7.91%, and follow-up 8.69%. Success of the intervention was, therefore, not attributable to an absolute increase in teacher attention to target children. A re-examination of Table 12.16, which depicts relative attention administered to target children and peers, also produced little evidence of a tracking effect. During intervention, groups E_3 and E_4 actually paid less relative attention to target children than did teachers in E_1 and E_2. This difference was quite small, however, and was not maintained through follow-up.

Several additional measures of teacher attention were also computed. Table 2.17 presents the percentage of total teacher responses to the target child which were disapproving or critical. As anticipated, a large decrease from baseline levels ($\bar{x}$ Phase I+III = 14.89%) was obtained during intervention for each experimental group ($\bar{x}$ = 8.20%) and this trend continued through follow-up ($\bar{x}$ E_1-E_4 = 5.55%). Group C_1 teachers failed to decrease their percentage of critical statements or gestures below 12% at any point. Concomitant with decreases in criticism were increases in the percentage of teacher responses which could be labeled as "praise." Table 2.18 shows that this percentage more than doubled during the intervention period for experimental groups while remaining stable for C_1. It appears that intervention was responsible for increasing positive comments and reducing negative ones without requiring any increase in either absolute or relative attention directed to the target child.

Teacher attention has also been examined with respect to "disruptiveness" and "distractibility." The results are consistent with those reported for total deviant behavior (minus "looking around"). Consequently, they will not be included among results relevant to these variables.

Disruptiveness (PA) + (DI) + (T-) + (P-)

Table 2.19 shows means and standard deviations of summed z scores for target children and peers. Figure 2.2 presents scores by observation for acting-out target children, peers, and matched controls in group C_1. The analysis of variance (Table 2.20) yielded significant main effects for target/peer (F = 6.297, df = 1, 38; p < .02) and phase (F = 3.114, df = 5, 190; p < .01). However, variability contributed by the group factor failed to reach

Table 2.17

Percentage of Teacher Responses to the Target Child Which were Disapprovals (DI)

	Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phases IV-V Intervention	Phase VI Follow-up
E_1	16.73	15.94	10.51	8.82	5.39
E_2	21.20	12.26	16.89	7.09	4.71
E_3	14.10	20.16	18.18	11.84	9.17
E_4	9.84	7.76	9.67	4.99	2.56
C_1	12.81	18.03	12.16	15.34	12.04

Table 2.18

Percentage of Teacher Responses to the Target Child Which were Approvals (AP)

	Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phases IV-V Intervention	Phase VI Follow-up
E_1	8.37	9.42	8.14	14.19	11.20
E_2	5.70	4.36	4.36	13.69	13.15
E_3	7.45	5.76	6.06	16.63	11.31
E_4	4.92	6.52	1.43	12.14	12.82
C_1	6.69	3.68	6.42	5.71	5.47

Table 2.19

Mean z Scores and Standard Deviations for Disruptiveness

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	8.10839	4.33020	4.73060	8.57000	4.26900	2.23080
	Peer	3.06740	-0.30760	2.99740	5.87599	1.09140	2.21900
E_2	Target	4.47580	6.59080	8.33899	2.49400	1.67020	1.87240
	Peer	-1.01940	1.58960	5.03540	0.92460	1.80520	2.01280
E_3	Target	12.16779	11.75559	7.35280	9.20919	6.32840	6.03739
	Peer	6.08039	5.87739	5.20080	6,36519	4.13920	4.57680
E_4	Target	4.41450	5.53475	4.60125	5.37975	4.61875	1.08900
	Peer	-1.32670	1.18850	1.93750	1.12500	2.09870	0.86900
C_1	Target	5.70080	6.22279	3.68560	8.71920	6.09859	7.95120
	Peer	4.73539	4.88959	1.11040	5.58960	4.76419	3.05980

Table 2.19 (continued)

Group		Standard deviations					
		Phase I Basline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	6.58394	3.33937	6.02233	4.29181	4.65214	4.56879
	Peer	3.50215	2.63944	4.07643	4.55114	2.56482	4.28659
E_2	Target	5.39532	6.29523	5.87674	6.32353	5.70125	6.23212
	Peer	0.62110	3.20162	2.10938	3.82572	1.82703	1.84861
E_3	Target	8.33664	7.44396	2.14997	7.94113	5.10926	5.94419
	Peer	5.82455	3.19205	1.51662	3.82503	5.82930	2.76459
E_4	Target	4.21490	4.74280	1.68304	3.98200	3.67311	4.36008
	Peer	2.98830	2.60650	1.78800	4.11130	3.57960	2.46110
C_1	Target	5.88632	6.93435	5.63438	4.58182	8.57866	5.23545
	Peer	6.39477	7.40872	3.57815	5.92206	4.70506	4.50379

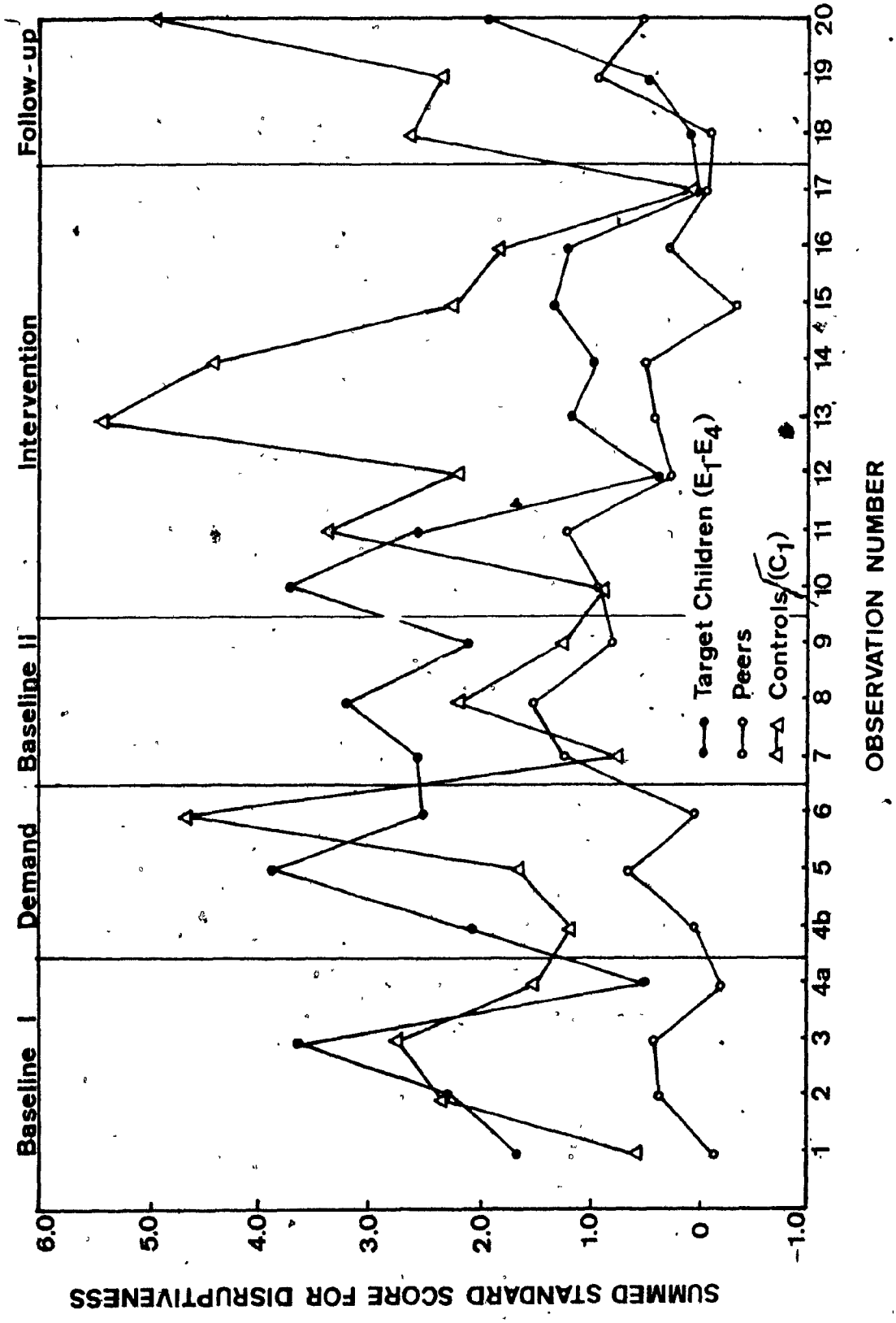


Figure 2.2. z Scores for Disruptiveness for Total Deviant Behavior (excluding LO) for Treated Target Children, Peers, and Matched Controls

Table 2.20

Analysis of Variance for Disruptiveness

Source	SS	DF	MS	F
Mean	5452.89844	1	5452.89844	57.19095
Target child/peer (A)	600.36841	1	600.36841	6.29677*
Group ($\hat{B}$)	757.88794	4	189.47198	1.98722
A x B	12.51880	4	3.12970	0.03282
Error	3623.12769	38	95.34546	
Phase (C)	153.03638	5	30.60727	3.11418**
A x C	105.39331	5	21.07866	2.14468
B x C	580.95825	20	29.04791	2.95552***
A x B x C	132.59741	20	6.62987	0.67457
Error	1867.38818	190	9.82836	

* $p < .05$

** $p < .01$

*** $p < .001$

significance (F = 1.987, df = 4, 38; p < .12). A significant interaction was obtained for phase x groups (F = 2.955, df = 20, 190; p < .001).

The mean for groups E_1-E_4 during natural baselines was 6.893 as compared to 7.133 during the demand condition. This difference was not significant (F = 0.102, df = 1, 190; n.s.), indicating no influence whatever due to the instructional set. Despite increased teacher attention to the subject, disruptive behaviors remained at high level.

By late intervention (Phase V), disruptive behavior (of E_1-E_4) decreased to 4.201 or approximately 2/3 of a standard deviation reduction for each of the four component behaviors. This improvement was significant (F = 9.091, df = 1, 190; p < .01). The control group did not exhibit any significant change as the baseline mean (Phases I and III) of 4.693 was actually lower than disruptiveness at late intervention ($\bar{x}$ = 6.099). In fact, controls continued to deteriorate through follow-up, where the mean disruptiveness score reached 7.951.

While a treatment effect was obtained across experimental groups, E_1 and E_4 did not demonstrate this until Phase VI. Therefore, the mean disruptiveness score during follow-up (2.897) was considerably lower than that obtained during later intervention (4.201). When one compares Phase VI to the baseline, the effects are quite dramatic. Disruptiveness for experimental subjects decreased an average of one standard deviation for each of the four component responses as a reduction from 6.893 to 2.897 was observed across the four response categories.

At follow-up, the experimental group mean of 2.897 compared very favorably to that of the control group, whose disruptiveness score was 7.951. This difference was highly significant (F = 10.635, df = 1, 190; p < .01), thus

showing that untreated, acting-out children differ substantially from those who have participated in an operant program.

Extremely encouraging was the discovery that the follow-up score for experimental target children ($\bar{x}$ = 2.897) did not differ significantly from normal peers ($\bar{x}$ = 2.531; F = 0.144, df = 1, 190; n.s.). The intervention was successful in bringing disruptive behaviors within normally accepted limits. On the other hand, the disruptive composite continued to discriminate control subjects from their peers (F = 6.086, df = 1, 190; $p < .05$) during the concluding phase of the study.

Orthogonal comparisons between experimental groups were also performed to test for differential effects of observation training, practice, and monitoring. While inter-group variation was found in a few cases, no clear pattern emerged which would indicate that decreased disruptiveness was a function of observational input.

Distractibility (HR) + (P-) + (IL) + (SS) + (LO) + (NA)

Means and standard deviations of summed scores for each group of target children and peers are presented in Table 2.21. Figure 2.3 shows z scores by observation for acting-out target children, peers, and matched controls. There was a substantial decrease in distractibility for each of the experimental groups and comparatively little improvement for the controls from baseline to intervention. Results of the ANOVA (see Table 2.22) showed main effects for target/peer (F = 30.453, df= 1, 38; $p < .001$) and phase (F = 26.322, df = 5, 190; $p < .001$). No main effect for group was obtained (F = 0.917 df = 1, 38; (n.s.). Again, orthogonal contrasts were performed in order to test specific hypotheses. The results were as follows.

Table 2.21

Mean z Scores and Standard Deviations for Distractibility

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	10.18199	10.55459	4.03660	7.10600	2.31180	0.07960
	Peer	5.69560	3.37360	1.26160	2.03060	-2.15680	-0.18300
E_2	Target	13.00359	5.43759	9.14419	5.94199	3.39320	2.93700
	Peer	3.64839	2.49520	2.95120	-1.82640	2.59220	0.83880
E_3	Target	11.85939	13.23739	6.56879	4.12700	2.42700	0.49600
	Peer	6.95040	4.97459	5.05739	1.18060	1.06100	2.51320
E_4	Target	13.69650	6.58699	4.09725	4.65175	2.43425	2.16925
	Peer	4.14270	0.26700	0.10450	-0.51200	0.08820	-0.44970
C_1	Target	12.17719	9.51539	4.89560	6.84039	8.11419	10.53559
	Peer	4.34739	3.41480	0.88920	-0.25020	2.09960	1.51060

Table 2.21 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	6.34366	2.36701	3.44984	3.54033	4.21488	3.29326
	Peer	2.72617	3.48653	1.40411	3.28494	3.60604	2.02447
E_2	Target	7.11389	5.53615	6.02749	4.15472	5.27152	7.69298
	Peer	2.32313	1.63987	3.89932	1.57737	3.25850	2.57139
E_3	Target	2.48506	4.02153	2.72815	4.19567	3.17765	3.52130
	Peer	4.61662	5.66289	3.12397	4.25971	3.65099	4.35285
E_4	Target	3.24075	1.86628	2.51842	5.54263	4.64458	8.29159
	Peer	3.17930	2.25630	4.03800	3.59940	1.80870	2.15630
C_1	Target	6.64269	3.93168	3.44392	5.38333	6.16055	5.94189
	Peer	4.50676	5.68943	2.85278	5.43066	5.32167	3.45596

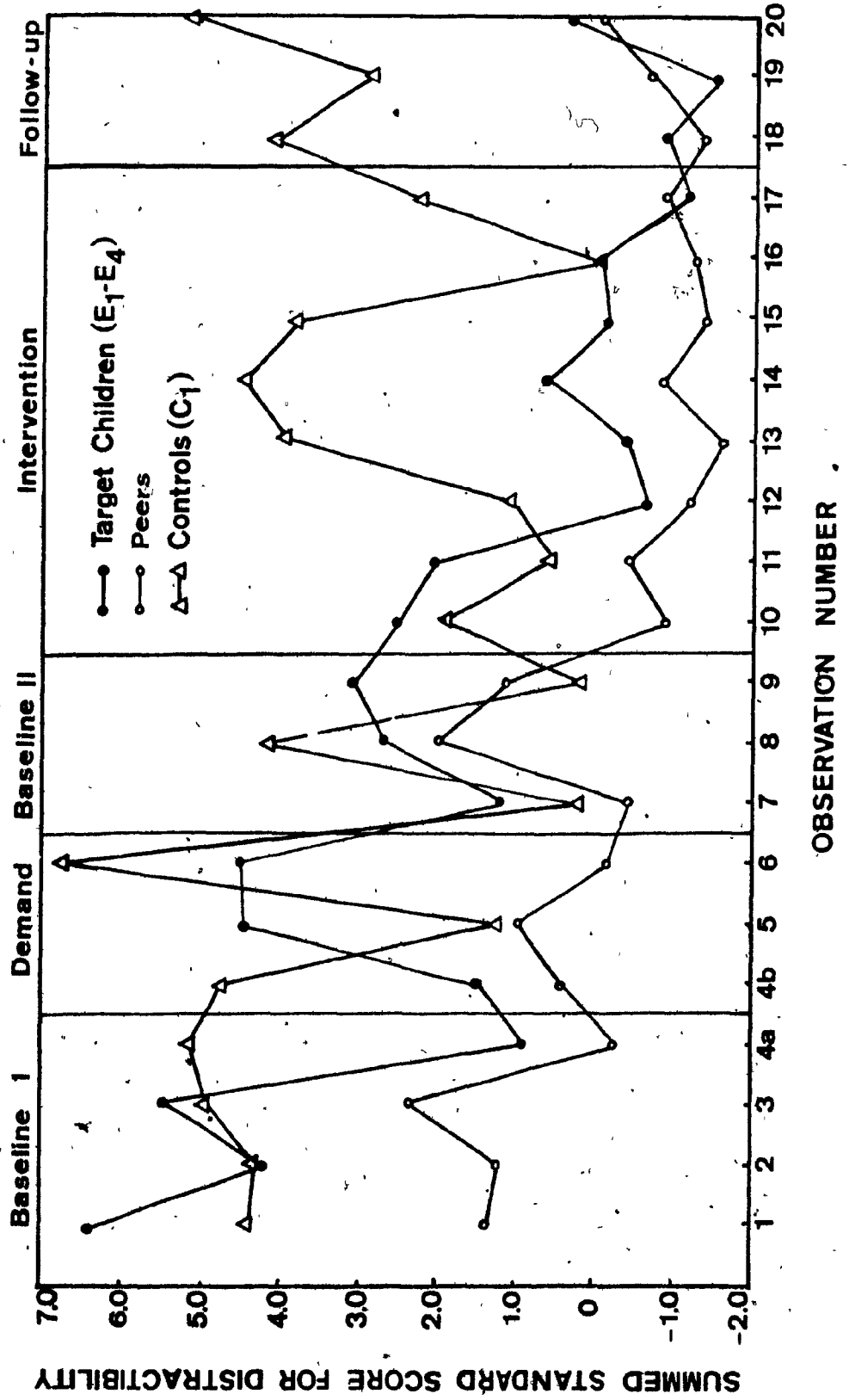


Figure 2.3. z Scores for Distractibility for Total Deviant Behavior (excluding LO) for Treated Target Children, Peers, and Matched Controls

Table 2.22

Analysis of Variance for Distractibility

Source	SS	DF	MS	F
Mean	5228.37891	1	5228.37891	101.70796
Target child/peer (A)	1565.48730	1	1565.48730	30.45351***
Group (B)	188.53906	4	47.13477	0.91692
A x B	119.03369	4	29.75842	0.57889
Error	1953.42065	38	51.40579	
Phase (C)	1545.60913	5	1545.60913	26.32172***
A x C	220.79565	5	44.15912	3.76015**
B x C	564.48901	20	28.22444	2.40331**
A x B x C	259.71680	20	12.98584	1.10574
Error	2231.35693	190	11.74398	

** $p < .01$

*** $p < .001$

As was the case with "total deviant" and "disruptiveness," the demand procedure was not effective in reducing the proportion of distractible behavior. The overall mean for experimental groups during natural baseline was 9.081, compared to 9.079 during the second phase. This difference was not significant ($F = 0.015$, $df = 1, 190$; n.s.).

The behavioral intervention was very successful in reducing distractibility. The natural baseline mean of 9.081 was reduced to 2.652 during late intervention, an average decrease of 1.07 standard deviations for each of the six component behaviors. This treatment effect was highly significant ($F = 44.206$; $df = 1, 190$; $p < .01$). Control subjects showed a far different pattern. The score for C_1 decreased substantially as baseline continued. The mean for the

three sets of baseline observations was 12.177, 9.515, and 4.895. However, as intervention began for E_1-E_4, distractibility rose for C_1 and maintained the increase through follow-up, when the mean was 10.535. Comparisons between Phases I-III vs. V yielded an insignificant F of 0.179 (df = 1, 190; n.s.), showing that mere passage of time and the normal educational routine were not effective in reducing distractibility.

Distractible behavior continued to decrease during Phase VI. The mean score for E_1-E_4 during this phase was only 1.945 or only about one-third of a standard deviation above the baseline peer norms for each of the six component behaviors.

At follow-up, the experimental subjects differed markedly from untreated controls, the latter showing a mean distractibility score of 10.535. This difference was significant (F = 27.950, df = 1, 190; p < .01) and indicates that the behavioral intervention did, indeed, account for the improvement shown by acting-out children.

Further examination of the follow-up data revealed that no significant difference existed between target subjects and peers (F = 0.440, df = 1, 190; n.s.). The control group target children, however, differed substantially from their peers. The score at follow-up for the former was 10.535, compared to only 1.511 for the latter (F = 17.339, df = 1, 190; p < .01).

Social Withdrawal (P+) + (VO) + (IT) - (SS) - (LO)

Results for this composite variable were not as clear-cut or easily interpretable as those obtained for acting-out children. Table 2.23 gives the phase means and standard deviations of this combination for each group of target children and peers. Unlike the previous analyses, where a decrease in

Table 2.23

Mean z Scores and Standard Deviations for Social Withdrawal

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	-8.58779	-2.76340	-4.23640	-3.17240	0.33180	0.30140
	Peer	-1.62760	2.12540	0.75100	1.99420	2.35620	3.55320
E_2	Target	-6.54059	-3.74240	-4.75499	-0.84080	-0.97240	-3.66120
	Peer	-0.30600	1.20700	1.03980	2.43460	1.86460	1.56500
E_3	Target	-5.07819	-1.89420	-4.45580	-0.77900	-2.80120	-1.36140
	Peer	-1.68620	0.04540	-2.58820	1.47560	1.24360	0.02720
E_4	Target	-7.40139	-3.80500	-3.72480	-4.03380	-2.32800	1.62500
	Peer	-0.67840	0.72300	0.00020	2.69860	2.07480	2.84260
C_1	Target	-6.97480	-5.24020	-3.90680	-1.72700	-2.76080	-2.86460
	Peer	0.14800	3.10980	1.17160	3.24080	4.23719	0.97660

Table 2.23 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	4.26371	2.97925	2.69272	5.51961	2.68480	4.41670
	Peer	2.31801	3.19399	2.64659	3.65821	1.64107	1.42161
E_2	Target	5.22613	1.85493	4.81684	5.28465	4.16302	4.32701
	Peer	3.57363	1.04420	1.97067	2.22333	2.05536	3.52999
E_3	Target	2.96820	3.19804	4.58105	3.75930	4.98088	3.46527
	Peer	0.77723	1.99525	2.18561	1.47799	3.35495	2.96297
E_4	Target	3.20290	1.64152	2.92770	6.47712	6.43901	3.76824
	Peer	1.29170	2.28050	0.58670	0.70910	2.56040	2.15220
C_1	Target	5.25930	5.07904	2.49263	7.82674	4.19840	5.69478
	Peer	3.07214	1.63958	0.70318	3.61215	2.32731	2.34462

summed z scores constituted improvement, social withdrawal became less pronounced as scores increased. The reader can readily detect a gradual rise in scores for all five groups. z scores by observation are presented graphically for withdrawn target children, peers, and matched controls in Figure 2.4.

When data were subjected to a 2 (target/peer) x 5 (group) x 6 (phase) analysis of variance for repeated measures, significant main effects were found for target/peer (F = 40.512, df = 1, 40; $p < .001$) and phase (F = 18.220, df = 5, 200; $p < .001$). However, no effect for groups was obtained (F = 0.126, df = 4, 40; n.s.); nor were any of the interaction terms significant (see Table 2.24). As hypotheses were formulated a priori, orthogonal comparisons between means were still conducted, but were restricted solely to the specific contrasts built into the design or suggested by the theoretical basis for the experiment (Winer, 1971).

A powerful effect was demonstrated for the high demand procedure (Phase II) when compared to both natural baselines across groups (F = 11.159, df = 1, 200; $p < .01$). Teachers in groups E_1-E_4 were able to raise composite z scores from a mean of -5.597 to -3.051 when instructed to "make the child appear outgoing." This was accomplished without providing children with joint tasks or otherwise deviating from the curriculum. The reader will recall that five types of activities were designated and monitored: group, individual, structured, unstructured, and transitional. The proportion of time in which each activity occurred is presented by phase for combined experimental groups (E_1-E_4) in Table 2.25. A 5 (activity) x 4 (group) x 5 (phase) analysis of variance for repeated measures was performed in order to test for the possibility that teachers altered the classroom routine in order to stimulate the social behavior of the target child. Table 2.26 presents the results of the

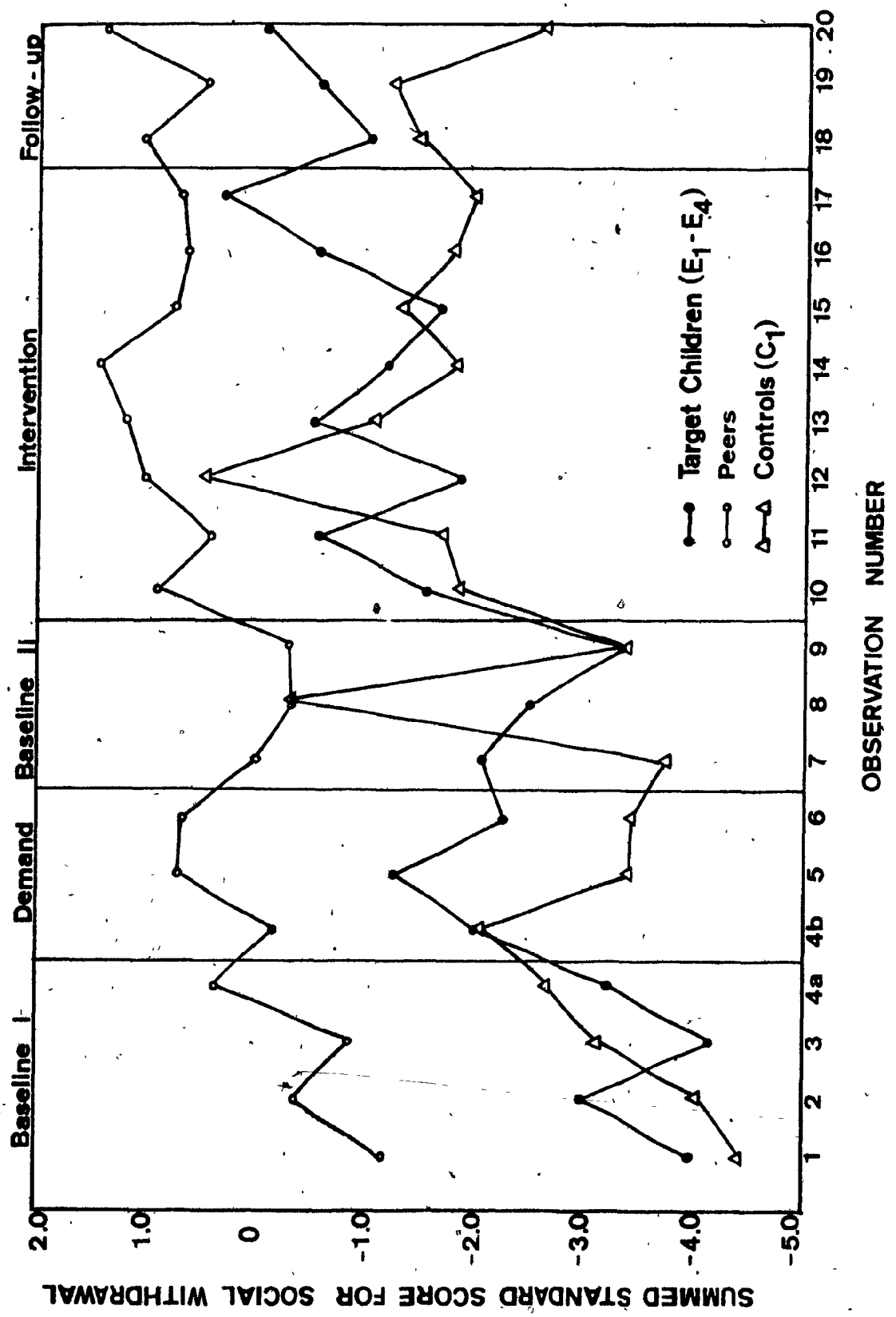


Figure 2.4. z Scores for Social Withdrawal for Total Deviant Behavior (excluding LO) for Treated Target Children, Peers, and Matched Controls

Table 2.24

Analysis of Variance for Social Withdrawal

Source	SS	DF	MS	F
Mean	321.69238	1	321.69238	8.68751
Target child/peer (A)	1500.13794	1	1500.13794	40.51221***
Group (B)	18.74292	4	4.68573	0.12654
A x B	98.36548	4	24.59137	0.66411
Error	1481.17163	40	37.02928	
Phase (C)	705.74438	5	141.14886	18.22025***
A x C	65.37329	5	13.07466	1.68775
B x C	193.36841	20	9.66842	1.24805
A x B x C	96.53271	20	4.82664	0.62305
Error	1549.36279	200	7.74681	

*** $p < .001$

Table 2.25

Proportion of Time in Which Each Activity Occurred for Withdrawn Children

	Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up	Marginal $\bar{x}$
Individual/ Unstructured	.020	.015	.006	.029	.012	.031	.018
Individual/ Structured	.450	.407	.462	.454	.484	.491	.458
Group/ Unstructured	.021	.019	.028	.041	.034	.016	.027
Group/ Structured	.441	.526	.453	.415	.406	.416	.443
Transitional	.069	.033	.051	.068	.064	.046	.055

Table 2.26

Analysis of Variance for Activity Proportions

Source	SS	DF	MS	F
Mean	322.20288	1	322.20288	1066.42847
Activity (A)	214.80724	4	53.70180	177.74243***
Error	28.70261	95	0.30213	
Phase (B)	0.26266	5	0.05253	0.58751
A x B	2.52835	20	0.12642	1.41382
Error	42.47244	475	0.08942	

*** $p < .001$

ANOVA on the arcsin transformed activity proportion scores. Absence of a main effect for Phase ($F = 0.587$, $df = 5, 475$; n.s.) and no significant activity x phase interaction ($F = 1.414$, $df = 20, 475$; n.s.) indicates that the distribution of activities did not vary as a function of the demand procedure. Hence, teachers did not deviate from their normal routine. However, the percentage of intervals in which the teacher responded to the target child increased from an experimental group mean of 4.63% in Phase I to 8.11% under the high demand condition (see Table 2.27). The ratio of teacher attention to the target child divided by her attention to peers rose from .892 to 1.237 (see Table 2.28). Certainly, the increases in both absolute time and relative proportion of attention directed toward the target child suggest that teachers followed the instructions by putting forth extra effort.

The fact that the withdrawal score for the control group showed an increase from -6.975 in Phase I to -5.240 in Phase II suggests that the demand procedure did not account for all of the changes in the experimental groups'

Table 2.27

Percentage of Intervals in Which Teacher Attended to the Withdrawn Target Child

	Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phases IV-V Intervention	Phase VI Follow-up
E_1	4.11	5.90	3.27	4.13	4.00
E_2	3.73	5.52	4.38	4.06	3.14
E_3	5.19	8.80	5.05	4.41	4.76
E_4	5.49	12.23	8.63	6.18	7.65
C_1	4.70	4.27	3.14	4.60	2.48
$\bar{x}$ (E_1-E_4)	4.63	8.11	5.33	4.69	4.89

Table 2.28

Relative Teacher Attention to Withdrawn Target Child

	Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phases IV-V Intervention	Phase IV Follow-up
E_1	.806	1.302	.421	.777	.433
E_2	.544	.925	.770	.721	.425
E_3	1.343	.951	1.300	.677	.785
E_4	.875	1.771	1.000	.847	.846
C_1	.826	.904	.532	.757	.385
$\bar{x}$ (E_1-E_4)	.892	1.237	.873	.755	.622

withdrawal scores. However, since the increase in social and attending behaviors shown by C_1 (1.735) was considerably less than that demonstrated by E_1-E_4 ($\bar{x}$ = 3.625), it is reasonable to conclude that the instruction was highly influential. A complete return to baseline was not evidenced in Phase III, although a trend in this direction was clearly demonstrated (see Figure 2.3).

When the demand baseline withdrawal levels of E_1-E_4 were compared with those observed during late intervention, a marginally significant effect was obtained (F = 3.341, df = 1, 200; p < .1). The means across experimental groups for these two phases were -3.051 and -1.442 respectively. While this difference approaches the conventional .05 significance level, the most parsimonious conclusion is that by late intervention, the training given teachers was not successful in raising social behavior above the level obtained during brief periods of optimal teacher motivation.[13] However, improvement persisted through follow-up for all groups except E_2 and C_1. The overall mean for experimental groups at follow-up rose to .774. Orthogonal comparisons showed that the means of the four experimental groups were significantly higher in Phase VI than in Phase II (F = 6.694, df = 1, 200; p < .02). In light of the fact that the control group showed no further progress from Phase V to VI, it is reasonable to conclude that the behavior modification techniques were at least partially successful in raising levels of appropriate social and attending behavior above those observed in either a natural or high demand condition.

---

[13] Since Phase V (late intervention) includes behavior samples up to four weeks prior to termination, this comparison might underestimate the eventual impact of treatment.

Use of the qualifier "partially" arises because of the substantial degree of improvement shown by untreated control subjects. The baseline mean across Phases I-III for C_1 was -5.374. This rose to -2.761 during Phase V and remained at the same level through follow-up (see Table 2.23). Orthogonal contrasts of the three baseline means with that of Phase V showed near-significant improvement ($F = 3.305$, $df = 1, 200$; $p < .1$), due either to maturation and/or variables associated with the normal educational process. The degree of improvement shown by C_1 was such that the group was rendered indistinguishable from E_1-E_4 at follow-up ($F = 2.257$, $df = 1, 200$; $p < .2$).

To summarize, the treatment groups showed considerably less withdrawn behavior when levels from either type of baseline are compared to scores at follow-up. Control students show somewhat less improvement from baseline but do progress to the point where they are statistically indistinguishable from treated cases at follow-up.

When comparisons between target children and peers were made, the effect of treatment becomes more complex. Table 2.23 gives the peer means and standard deviations along with those for target subjects. The Phase VI means across experimental groups for subjects and peers were .774 and 1.997, respectively. This difference was significant ($F = 9.912$, $df = 1, 200$; $p < .01$) and showed that target children did not attain scores in the normal range. However, closer examination of the data revealed that this difference was chiefly attributable to the failure of E_2 to rise to peer standards ($F = 8.814$, $df = 1, 200$; $p < .01$). Contrasts conducted for the other three treatment groups demonstrated that none of these differed significantly from peer norms at follow-up. F ratios were 3.412, 0.622, and 0.479 for E_1, E_3, and E_4, respectively. The fact that E_1 and E_2 differed from their peers more

than did E_3 or E_4 is some evidence of a differential tracking effect. This will be discussed further in subsequent paragraphs. The control group target children still differed from their peers at follow-up (F = 4.762, df = 1, 200; $p < .05$).

Were it not for the deterioration shown by E_2 at follow-up, one would be in a much stronger position to advocate use of the treatment package. In order to learn why this group did not fit the pattern displayed by the other three groups receiving training, separate analyses of the five component behaviors were conducted. Tables 2.29, 2.31, 2.33, 2.35, and 2.37 present the means and standard deviations (in z scores) for each of the categories P+, VO, IT, LO, and SS. Tables 2.30, 2.32, 2.34, 2.36, and 2.38 show the results of the 2 (subject/peer) x 5 (group) x 6 (phase) analysis of variance for each behavior. Without going into great detail, several interesting findings will be highlighted. These shed light on the specific vs. general nature of the treatment effects, the discrepant pattern shown by E_2 at follow-up, and the presence of a (group) tracking effect.

The behavior "appropriate interaction with peers" (P+) clearly discriminated target children from their peers during baseline (t = 5.642, df = 24; $p < .01$). This can be seen as the most clinically relevant of the five responses forming the withdrawal composite. Treatment techniques dealing with social interaction were heavily emphasized and their implementation superceded work on other behaviors. Only when peer contact rose to a level considered satisfactory to the teacher were other problems treated directly. Because this particular category of behavior is considered prosocial, a rise in z score indicates increasing interaction with peers. Table 2.29 shows that all five groups improve from baseline to treatment, and that this change is

Table 2.29

Mean z Scores and Standard Deviations for Appropriate Interaction With Peer

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	-1.34960	0.21480	0.10260	0.91960	1.81000	2.38000
	Peer	-0.43080	0.68120	0.87860	1.06120	1.79160	1.63280
E_2	Target	0.80620	0.97420	0.71960	1.39460	1.54220	1.14500
	Peer	0.90440	0.89660	1.83900	2.42460	1.66320	1.85280
E_3	Target	-1.49740	-0.51680	-1.38740	0.89060	0.63520	-0.02680
	Peer	-0.49000	-0.59080	-0.75380	0.64760	0.75720	0.08900
E_4	Target	-1.40900	-0.31500	-0.70980	0.69000	0.71400	0.49320
	Peer	-0.08120	0.98180	0.91340	2.33560	2.21620	1.15540
C_1	Target	-0.87280	-0.99460	-0.32700	-0.47580	-0.41980	-0.47700
	Peer	0.67240	0.02020	0.75420	0.71900	0.77880	0.66500

Table 2.29 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	1.08697	0.94247	0.60479	1.98326	0.75197	1:09608
	Peer	0.72902	0.91108	0.80294	0.87498	0.68271	0.93722
E_2	Target	2.52229	1.37618	1.19254	1.78854	1.35240	1.47619
	Peer	2.32496	1.70900	0.66883	1.26649	0.66121	1.37209
E_3	Target	0.37267	1.66336	1.05619	1.15904	1.66295	1.24912
	Peer	0.39555	2.17792	0.75305	0.90428	1.42184	0.63663
E_4	Target	1.47893	1.11260	1.46229	1.65147	2.13500	1.45002
	Peer	0.96810	1.41190	1.35310	0.70010	0.88530	0.72310
C_1	Target	1.09140	0.56395	0.48639	0.47440	0.84062	1.25521
	Peer	1.03860	1.21990	0.49128	0.64628	0.58966	0.84686

Table 2.30

Analysis of Variance for Appropriate Interaction with Peer

Source	SS	DF	MS	F
Mean	78.22421	1	78.22421	28.11015
Target child/peer (A)	37.92183	1	37.92183	13.62735***
Group (B)	92.18100	4	23.04524	8.28139***
A x B	16.33899	4	4.08475	1.46787
Error	111.31094	40	2.78277	
Phase (C)	93.70612	5	18.74121	15.00202***
A x C	4.36345	5	0.87269	0.69857
B x C	44.34308	20	2.21715	1.77479*
A x B x C	8.46027	20	0.42301	0.33861
Error	249.84930	200	1.24925	

* $p < .05$

** $p < .01$

*** $p < .001$

Table 2.31

Mean z Scores and Standard Deviations for Volunteering

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	-1.05780	-1.17580	-1.06620	-0.50580	-0.83340	-1.04120
	Peer	0.59960	0.33780	-0.16820	0.37800	0.32660	0.06080
E_2	Target	-0.55560	-0.69440	-0.79100	0.66820	-0.46640	-1.26540
	Peer	0.62140	0.42680	0.30280	0.11240	0.12420	-0.43260
E_3	Target	0.57900	0.59300	0.67380	-0.23400	-0.39160	-0.05140
	Peer	0.81100	1.29440	0.79340	0.45920	0.88840	0.45060
E_4	Target	-1.22600	-0.85800	-0.48820	-1.23520	-0.61620	-0.69800
	Peer	0.33360	0.53400	0.33520	0.27000	0.51900	0.30960
C_1	Target	-0.48240	-0.07440	-0.17680	0.71640	-0.76660	-1.26820
	Peer	0.80380	1.06900	0.82140	1.47720	1.15900	0.28760

Table 2.31 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	0.88867	0.52864	0.58391	1.89825	0.67341	1.04886
	Peer	0.77036	0.94683	0.57313	1.96866	1.60739	1.19598
E_2	Target	1.29874	1.07550	1.55535	1.21717	1.34587	0.66811
	Peer	2.70079	1.84642	1.96498	1.89431	1.21891	1.79093
E_3	Target	1.03982	1.48415	1.51912	0.63839	0.46122	0.80978
	Peer	0.73375	1.07357	0.46145	0.71708	0.66484	1.03825
E_4	Target	1.21230	1.05581	0.96793	1.25983	1.07598	0.55441
	Peer	1.07530	0.84860	0.92700	0.93280	0.50060	0.87500
C_1	Target	0.63031	1.02320	0.71238	0.81562	0.69064	0.75699
	Peer	0.47384	1.06083	0.37936	0.95483	0.38009	0.93639

Table 2.32

Analysis of Variance for Volunteering

Source	SS	DF	MS	F
Mean	0.05602	1	0.05602	0.01758
Target child/peer (A)	82.32965	1	82.32965	25.83330***
Group (B)	34.55246	4	8.63811	2.71046*
A x B	5.06348	4	1.26587	0.39720
Error	127.47836	40	3.18696	
Phase (C)	8.07028	5	1.61406	1.77960
A x C	1.86061	5	0.37212	0.41029
B x C	19.44513	20	0.97226	1.07197
A x B x C	5.46950	20	0.27347	0.30152
Error	181.39542	200	0.90698	

* $p < .05$

** $p < .01$

*** $p < .001$

Table 2.33

Mean z Scores and Standard Deviations for Initiation to Teacher

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	-2.07300	-0.69100	-1.09160	-1.59820	-1.15460	-1.50240
	Peer	0.39060	0.36040	0.76380	1.09140	0.59360	0.18300
E_2	Target	-1.85520	-1.72240	-0.65980	-0.55960	-1.00680	-1.44780
	Peer	-0.31520	-0.19300	0.24800	0.07600	-0.02200	0.03840
E_3	Target	-0.79240	-0.63340	-0.78880	-0.11820	-0.99560	-0.71080
	Peer	-0.62980	-0.34280	-0.28340	0.48720	-0.09100	0.72720
E_4	Target	-1.09120	-0.04960	-0.89200	-0.23940	-1.07700	0.95340
	Peer	0.74480	0.92680	-0.12780	1.13200	0.22240	1.19720
C_1	Target	-0.00920	-0.92000	-0.51680	0.01980	0.33060	-1.07220
	Peer	0.10080	0.76320	0.46300	0.66440	0.76520	-0.09180

Table 2.33 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	2.11118	1.80306	1.45658	1.44938	1.84073	1.30604
	Peer	1.17371	1.37190	1.12724	1.48391	2.05447	0.64563
E_2	Target	0.37979	0.54477	0.47828	0.29254	1.45742	1.39070
	Peer	0.88663	1.02565	0.68036	0.51431	0.77329	0.55338
E_3	Target	1.42871	1.29935	1.13023	1.47763	0.82567	1.22707
	Peer	0.49816	0.44633	0.88361	1.08018	0.93754	0.99266
E_4	Target	0.57809	1.17623	0.49429	2.57622	0.97422	1.40275
	Peer	0.99130	0.77080	0.75680	1.83150	0.89690	1.70720
C_1	Target	2.12928	0.87210	1.03357	2.66983	2.35569	0.97099
	Peer	1.16374	0.82462	1.32943	1.25900	2.06817	1.11319

Table 2.34

Analysis of Variance for Initiation to Teacher

Source	SS	DF	MS	F
Mean	16.62030	1	16.62030	3.51248
Target child/peer (A)	95.24590	1	95.24590	20.12895***
Group (B)	24.18768	4	6.04692	1.27794
A x B	14.34993	4	3.58748	0.75817
Error	189.27144	40	4.73179	
Phase (C)	10.87157	5	2.17431	1.99033
A x C	0.41591	5	0.08318	0.07614
B x C	32.74640	20	1.63732	1.49878
A x B x C	15.96614	20	0.79831	0.73076
Error	218.48755	200	1.09244	

*** $p < .001$

Table 2.35

Mean z Scores and Standard Deviations for Looking Around

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	2.76140	1.20480	1.83960	0.97580	-0.03680	-0.66520
	Peer	1.22180	-0.38680	0.23200	0.15240	-0.36200	-1.42460
E_2	Target	3.42420	1.27880	1.08240	0.64740	0.75660	0.77860
	Peer	1.10780	0.13960	0.33080	-0.1140	0.03700	-0.35940
E_3	Target	2.66440	0.31840	1.08120	-0.16040	0.49820	-0.38660
	Peer	0.97740	-0.17560	0.97880	-0.43220	-0.23460	0.21940
E_4	Target	1.85460	0.79320	0.35300	0.95600	-0.12840	-0.60420
	Peer	0.74320	0.67300	0.32280	0.25820	0.08780	0.06940
C_1	Target	2.73680	1.38820	1.00860	0.14920	-0.09040	-1.18140
	Peer	0.71900	-1.03620	0.16720	-1.15680	-1.47220	-0.71480

Table 2.35 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	1.73631	1.38967	0.76898	1.18832	1.54487	1.79852
	Peer	1.21055	1.87786	0.80504	0.53017	0.58289	0.68384
E_2	Target	2.62063	0.92084	1.11020	2.62868	1.97631	3.09396
	Peer	1.20402	0.68630	0.96097	0.80215	1.33018	1.63108
E_3	Target	1.59677	0.93030	1.65638	1.49864	1.62566	1.04948
	Peer	0.88875	0.51791	1.09626	0.37666	1.10813	1.71512
E_4	Target	1.82682	1.14316	0.99740	2.28266	1.43478	1.52565
	Peer	1.12150	0.70840	1.08610	1.01760	0.64200	0.90220
C_1	Target	1.80276	1.44951	0.69275	2.52605	0.84882	0.95933
	Peer	1.33807	1.28813	1.10388	1.37871	0.31870	0.82865

Table 2.36

Analysis of Variance for Looking Around

Source	SS	DF	MS	F
Mean	55.81369	1	55.81369	10.49142
Target child/peer (A)	50.90573	1	50.90573	9.56886**
Group (B)	15.61244	4	3.90311	0.73368
A x B	13.78107	4	3.44527	0.64761
Error	212.79739	40	5.31993	
Phase (C)	156.54671	5	31.30934	25.67836***
A x C	20.63582	5	4.12716	3.38489**
B x C	26.70247	20	1.33512	1.09500
A x B x C	12.81306	20	0.64065	0.52543
Error	243.85783	200	1.21929	

* $p < .05$

** $p < .01$

*** $p < .001$

Table 2.37

Mean z Scores and Standard Deviations for Self-Stimulation

Group		Means					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	1.34520	-0.09340	0.34100	1.01220	-0.47320	0.20040
	Peer	0.96520	-0.36000	0.49100	0.38400	0.71760	-0.25240
E_2	Target	1.51180	1.02120	2.94160	0.36020	0.28420	1.31460
	Peer	0.40920	-0.21560	1.01940	0.27900	-0.13620	0.25320
E_3	Target	0.70260	1.01880	1.87260	1.47740	1.55120	0.95980
	Peer	0.39980	0.49060	1.36520	0.55100	0.54540	1.01980
E_4	Target	1.82020	1.78940	1.28180	2.29300	1.47700	-0.27240
	Peer	0.93240	1.04640	0.79740	0.78120	0.79520	-0.24980
C_1	Target	2.87360	1.86340	1.87740	1.83780	1.99520	1.22860
	Peer	0.71020	-0.22140	0.69980	0.77660	-0.06220	0.59900

Table 2.37 (continued)

Group		Standard deviations					
		Phase I Baseline I	Phase II Demand	Phase III Baseline II	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	Target	1.07004	1.59194	0.88461	1.59300	0.63033	1.21803
	Peer	0.78626	1.24067	0.78164	1.06854	0.80955	0.69224
E_2	Target	1.93443	1.50324	2.33563	1.76435	1.46229	1.85276
	Peer	1.27984	1.10883	0.53984	0.46354	0.77099	0.98923
E_3	Target	1.04637	0.83159	1.34493	1.22967	0.98272	0.99882
	Peer	0.82153	0.96969	0.93170	0.58859	0.64806	1.08521
E_4	Target	1.26109	1.08653	0.86214	3.02899	2.37554	1.22858
	Peer	1.36080	1.27030	0.25750	0.50420	1.36530	0.76590
C_1	Target	1.57217	2.64281	1.53538	2.38900	2.67979	2.81264
	Peer	0.55026	0.39282	0.63406	0.80508	0.51404	0.66702

Table 2.38

Analysis of Variance for Self-Stimulation

Source	SS	DF	MS	F
Mean	224.84816	1	224.84816	50.87912
Target child/peer (A)	43.63214	1	43.63214	9.87317**
Group (B)	25.16222	4	6.29055	1.42344
A x B	17.58246	4	4.39561	0.99465
Error	176.77049	40	4.41926	
Phase (C)	25.30737	5	5.06147	3.79823**
A x C	3.04370	5	0.60874	0.45681
B x C	45.90845	20	2.29542	1.72253*
A x B x C	18.24463	20	0.91223	0.68456
Error	266.51782	200	1.33259	

* $p < .05$

** $p < .01$

maintained through the follow-up period. Closer examination reveals that target children in Group E_2 actually interacted above the overall baseline mean (of zero) prior to treatment, thus making it difficult and unwarranted to increase the level.[14] Despite the high baseline means, E_2 still showed an increase in peer interaction. This high point was attained during the fifth phase and dropped off somewhat thereafter.

Using P+ as a criterion variable, it was found that treated groups improved substantially. The natural baseline mean was -.590; the demand condition mean was .089. When high demand scores were contrasted with those of late intervention ($\bar{x}$ = 1.175), a significant treatment effect was obtained (F = 9.442, df = 1, 200; $p < .01$). At follow-up there was no difference between target children and their peers (F = .272, df = 1, 200; n.s.) Treated groups differed significantly from C_1 (F = 6.965, df = 1, 200; $p < .01$). Also, control subjects did not show a great deal of improvement over time. When their baseline means were contrasted with those of Phase V, the resultant F ratio was 0.292 (df = 1, 200; n.s.). Therefore, the intervention was particularly effective in improving the proportion of time during which withdrawn children interacted in an appropriate manner. By the termination of treatment, they were experiencing a normal amount of peer contact. The same cannot be said for control subjects, who failed to improve significantly, and who were still interacting at a level one standard deviation below their peers. Figure 2.5 shows graphically the intra-group, inter-group, and normative comparisons.

---

[14] Group E_2 did show baseline (Phase I and III) deficits in volunteering (-.664) and initiating to teacher (-1.303). It also produced above normal proportions of self-stimulation (2.173) and the highest degree of looking around (2.343).

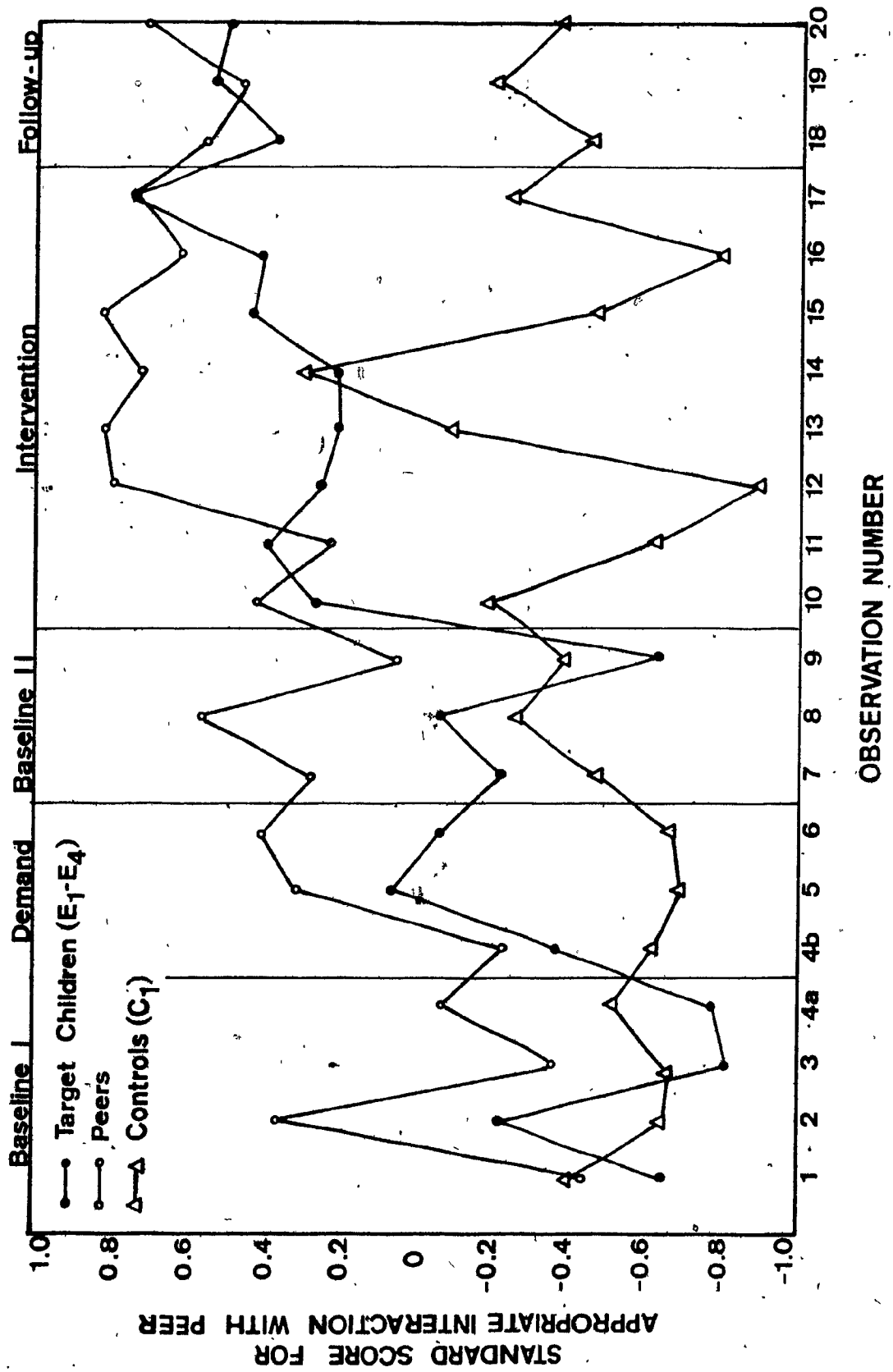


Figure 2.5. z Scores for Appropriate Interaction with Peer for Treated Target Children, Peers, and Matched Controls

The two other prosocial behaviors included in the withdrawal composite, "volunteering" and "initiation to teacher," showed no clear improvement for any of the five groups. Analysis of volunteering yielded non-significant main effects for phase (F = 1.780, df = 5, 200; p < .12) or any of the interactions involving this factor. This finding was identical for initiation to teacher. Again, the phase factor just failed to account for a significant proportion of variance (F = 1.990, df = 5, 200; p < .1). Interaction terms also failed to attain significance.

On the other hand, "looking around" and "self-stimulation" showed substantial changes in mean level. Table 2.35 shows a steady decrease in looking around across groups, including C_1. The ANOVA yielded a highly significant main effect for phase (F = 25.678, df = 5, 200; p < .001), but failed to produce a significant group effect (F = 0.734, df = 4, 40; p > .5). Thus, the most parsimonious conclusion is that socially withdrawn children, regardless of whether they receive treatment of the sort applied in the present study, show substantial reduction in looking around or daydreaming as a function of time and/or the normal educational process.

Analysis of "self-stimulation" revealed essentially the same pattern of results. Improvement over time was evidenced in a main effect for phase (F = 3.798, df = 5, 200; p < .003) but was not specific to any particular group(s) (F = 1.423, df = 4, 40; p < .25). A perusal of Table 2.37 shows that all groups decreased self-stimulation by at least one standard deviation by late intervention, with the exception of E_3. Interestingly, both E_1 and E_2 showed deterioration from Phase V to VI, the latter group increasing self-stimulation a whole standard deviation, from .284 to 1.315. This precipitous

change accounts for about half of the decline in the withdrawal composite score for E_2.

The amount of teacher attention to withdrawn children did not show much change from baseline through follow-up. The proportion of intervals in which there was contact between target child and teacher is presented in Table 2.27 for each group and phase. Application of behavior modification programs did not require an inordinate quantity of teacher time, as the average number of minutes actually dropped from 2.83 at baseline (Phase I and III) to 2.66 during intervention. Relative attention directed to target children (vis-a-vis peers) showed a similar pattern. During Phases I and III, target children received 88.24% of the teacher contact experienced by peers, while they received only 75.55% as much attention during intervention. Clearly, then, implementation of the program did not require teachers to significantly alter the distribution of their attention among students. This cannot be said for the demand procedures which, though effective, required a 46% increase in teacher attention from Phase I to III.

A search for a tracking effect yielded only one finding of psychological importance. Table 2.23 shows that groups E_1 and E_2 showed increased withdrawal from late intervention to follow-up. Certainly, the decline of .332 to .301 for E_1 is hardly noteworthy except in contrast to the substantial improvement demonstrated by both E_3 and E_4. Group E_3 showed an increased total z score by 1.44 in follow-up, while E_4 raised its score .703. Re-examination of Table 2.28 shows that relative teacher attention directed to target children was maintained from intervention to follow-up for E_3 and E_4, but fell dramatically for E_1 and E_2. Target children were receiving only about 43% of the teacher attention that their peers were in follow-up. This level was far below

those of baseline and intervention. Appropriate interaction with peers remained stable (see Table 2.29) but large increases in self-stimulation were evidenced. E_1 increased their proportion of self-stimulation by .67 standard deviation, while E_2 displayed a similar rise of 1.03 standard deviations. It would seem that peer-reinforced behavior is maintained in the absence of tracking. However, the teacher who is collecting data may be in an advantageous position to interrupt self-directed behavior and orient the child toward more appropriate activities. No additional benefits due to the external monitoring procedure were apparent in group E_4.

Walker Problem Behavior Identification Checklist

All scaled scores and total scores for both acting-out and socially withdrawn children were analyzed. However, only the data for the acting-out and distractibility scales will be presented, as these were most relevant to the disruptive/distractible subsample. Pre- and post-treatment means and standard deviations for the two factor scores and total score[15] are presented in Table 2.39, along with norms provided by Walker (1970).[16] Of the 25 children for

---

[15] Total score is derived by summing the scores for each of the five scales: acting out, withdrawal, distractibility, immaturity, and disturbed peer relations.

[16] The normative sample consisted of 534 pupils from grades 4-6. An identification sample of 1037 children from grades 1-3 was also obtained (Walker, 1971). The total score mean and standard deviation were 4.74 and 6.66 respectively for the latter group. Factor scores were not reported for the younger children and these data were subsequently lost (Walker, personal communication). Sampling bias may have invalidated the identification group as truly representative. However, their total score mean and standard deviation was substantially below that of the older normative sample (critical ratio 6.16, $p < .001$). The present sample may, in fact, have been relatively more deviant than the older age norms would indicate.

Table 2.39

WPBIC Scale and Total Scores for Acting-Out Target Children

Group		Acting-out		Distractibility		Total	
		Pre	Post	Pre	Post	Pre	Post
E_1	$\bar{x}$	10.40	4.40	10.20	3.80	31.40	11.20
	S.D.	8.20	4.10	.84	3.35	13.67	8.38
E_2	$\bar{x}$	15.00	8.00	9.80	5.60	37.00	18.00
	S.D.	7.31	6.04	1.92	4.22	15.15	10.25
E_3	$\bar{x}$	4.80	2.40	6.40	2.00	21.20	6.20
	S.D.	5.07	2.80	3.05	2.34	9.55	4.44
E_4	$\bar{x}$	6.00	5.25	7.00	4.00	18.50	11.20
	S.D.	1.83	5.50	2.45	4.55	6.24	11.93
C_1	$\bar{x}$	7.80	4.60	9.00	5.80	22.40	15.60
	S.D.	5.89	4.16	2.92	3.56	10.26	10.19
(E_1-E_4)	$\bar{x}$	9.10	5.00	8.42	3.84	27.50	11.70
	S.D.	6.62	4.98	2.74	3.65	14.19	9.57
Normative sample (Walker, 1970)	$\bar{x}$	2.23		2.63		7.76	
	S.D.	4.79		3.31		10.53	

whom pre-treatment data were obtained, all scored above the normal mean on one of the two scales. Fourteen children scored higher than one standard deviation above the mean on both scales; eight scored at this level only on the distractibility scale, and three others attained this level only on the acting-out scale. Two subjects did not appear at all deviant on either measure.

Acting-out. One can readily see that substantial decreases in teacher ratings of acting-out behavior were obtained following treatment.[17] Following intervention, the mean score for treated subjects was 5.01, the equivalent of a 44.6% reduction. Teachers also rated control subjects as improved. In fact, the post-treatment mean of 4.6 for C_1 was slightly lower than the average for treated subjects. The percentage reduction for C_1 was about the same, 41%. Scaled scores were subjected to a 5 (group) x 2 (occasions) ANOVA for repeated measures (see Table 2.40). A significant main effect for the occasion factor (B) was obtained (F = 13.370, df = 1, 19; p < .002), indicating that disruptive-distractible children improved over time. Failure to obtain a significant F ratio for the interaction indicates that this improvement was not confined to the experimental groups. Because untreated control children were perceived as acting out less at the completion of the study, the null hypothesis was not rejected; in other words, perceived change in acting-out behaviors cannot be considered a function of behavior modification training. Furthermore, there were no differences between experimental groups which could be attributed to a tracking effect.

---

[17] Consistent results on both the acting-out and distractibility scales were expected as inter-correlations between the two factors were .67 for the older normative sample and .49 for the grade 1-3 identification sample (Walker, 1971).

Table 2.40

Analysis of Variance for WPBIC Acting-Out Scores

Source	SS	DF	MS	F
Mean	2244.19800	1	2244.19800	48.13847
Group (A)	340.88940	4	85.22235	1.82804
Error	885.77319	19	46.61963	
Occasion (B)	178.29536	1	178.29536	13.36999**
A x B	61.62373	4	15.40593	1.15526
Error	253.37431	19	13.33549	

** $p < .01$

Distractibility. Table 2.39 gives the means and standard deviations for acting-out children on the distractibility scale. Once again, all five groups showed a decrease in score from the first rating period to the second. Groups E_1-E_4 produced an overall pre-treatment mean of 8.42 compared to 9.00 for C_1. Following intervention, the mean score for experimental children fell to 3.84, the equivalent of a 54% decrease in distractibility. Scores for control subjects decreased about 35%, to a mean of 5.8. The analysis of variance findings (see Table 2.41) showed a significant main effect for the occasion factor ($F = 39.010$, $df = 1, 19$; $p < .001$), but neither the group factor ($F = 1.599$, $df = 4, 19$; n.s.) nor the interaction term ($F = 0.798$, $df = 4, 19$; n.s.) accounted for significant variance. As was the case on the acting-out scale, both experimental and control children improved. However, examination of the factor norms showed treated cases scored within one standard deviation of the mean for Walker's normative sample, while scores for children in C_1 exceeded

this standard. Again, differences between experimental group means do not indicate a differential effect due to observation training, daily data collection, or external monitoring.

Table 2.41

Analysis of Variance for WPBIC Distractibility Scores

Source	SS	DF	MS	F
Mean	1926.16846	1	1926.16846	144.31139
Group (A)	85.37796	4	21.34448	1.59916
Error	253.59885	19	13.34731	
Occasion (B)	214.01749	1	214.01749	39.09940***
A x B	17.47890	4	4.36972	0.79832
Error	103.99985	19	5.47368	

*** $p < .001$

Social withdrawal. Means and standard deviations on the social withdrawal scale are presented in Table 2.42 for the subsample of children for which this measure was relevant. Also included are norms derived from ratings of pupils in grades 4-6. (Again, the reader is reminded that these norms are not age-appropriate to the present sample.) All 25 withdrawn target children scored at least one standard deviation above the norm while 18 of these were at least two standard deviations above it. The overall mean for experimental groups prior to their receiving training was 10.5. This was somewhat higher than the control group mean of 8.4. In fact C_1 ranked fifth, or the least withdrawn, among the five groups. (Newman-Keuls' tests showed that the mean for C_1 did not differ significantly from that of any other group ($df = 29$; $p > .05$ on all

Table 2.42

WPBIC Scale and Total Scores for Withdrawn Target Children

Group		Withdrawn		Total	
		Pre	Post	Pre	Post
E_1	$\bar{x}$	11.40	1.20	20.60	3.40
	S.D.	3.58	2.68	7.02	4.77
E_2	$\bar{x}$	9.60	2.20	26.60	7.40
	S.D.	1.52	2.95	7.13	2.70
E_3	$\bar{x}$	11.20	0.00	20.60	0.40
	S.D.	3.03	0.00	3.36	.89
E_4	$\bar{x}$	9.80	2.20	18.20	4.00
	S.D.	2.95	2.83	7.39	3.81
C_1	$\bar{x}$	8.40	8.80	19.20	28.00
	S.D.	1.52	5.21	14.96	26.73
(E_1-E_4)	$\bar{x}$	10.50	1.40	21.50	3.80
	S.D.	2.76	2.39	6.72	3.98
Normative sample (Walker, 1970)	$\bar{x}$	1.60		7.76	
	S.D.	3.19		10.53	

comparisons). Children in experimental groups showed very dramatic decreases in teacher-rated withdrawal. The post-treatment mean for E_1-E_4 was only 1.40, an 87% decrease from the first administration of the checklist. This compared very favorably to the control group, which actually scored higher during the second rating period ($\bar{x}$ = 8.80) than the first. Results of the ANOVA are presented in Table 2.43. As visual inspection of the means and standard deviations would suggest, highly significant F ratios were obtained for the occasion factor (F = 88.304, df = 1, 20; p < .001) and the interaction term (F = 51.969, df = 4, 20; p < .001). Newman-Keuls' tests between post-treatment means for each group showed that all treatment groups scored significantly lower than C_1 (df = 29; p < .01 on all comparisons). Therefore, substantial improvement was perceived by teachers in the training groups, whereas no progress whatever was reflected in teacher ratings of children in C_1.

Table 2.43

Analysis of Variance for WBPIC Withdrawal Scores

Source	SS	DF	MS	F
Mean	2086.57739	1	2086.57739	211.62071
Group (A)	59.71945	4	14.92986	1.51419
Error	197.19974	20	9.85999	
Occasion (B)	655.21777	1	655.21777	88.30412***
A x B	207.[illegible]807	4	51.96951	7.00397***
Error	148.40025	20	7.42001	

*** p < .001

Summary Reports

Eighteen weekly ratings on the seven-point scales of distractibility, disruptiveness, and social withdrawal were reduced to four means per subject on each variable. These represented the average summary report for baseline, early intervention, late intervention, and follow-up. A separate 4 (group) x 4 (phase) ANOVA was performed on each of the three variables. Since the control group submitted only two summary reports (pre and post), these were compared to the ratings from the initial and final weeks for experimental subjects.[18] A 5 (group) x 2 (occasion) repeated measures ANOVA was also conducted for each of the three variables.

Disruptiveness. The mean baseline rating across experimental groups was 5.49 (see Table 2.44). This indicates that subjects were considered quite disruptive prior to teacher training. A considerable decrease for the subsequent three phases was evidenced with the lowest point occurring at follow-up ($\bar{x}$ = 2.79). The first ANOVA (see Table 2.45) yielded a significant main effect for phase (F = 61.923, df = 3, 45; $p < .01$). Baseline levels of disruptiveness were substantially reduced over time. Failure to obtain either a significant main effect for groups (F = 1.004, df = 3, 15; $p > .4$) or a group x phase interaction (F = 0.978, df = 9, 45; $p > .4$) suggests that there was not a differential tracking effect. Low standard deviations show that teachers in each of the four experimental groups tended to rate children in much the same

---

[18] The effects of repeated readministration of the summary report scales to experimental teachers cannot be ascertained from the present study. The reader is warned that if the "practice" effect were large, any comparison to the ratings supplied by control teachers would be seriously confounded.

Table 2.44

Experimental Group Ratings of Distractibility

Group		Phases I and III Baseline	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	$\bar{x}$	5.30000	4.05000	3.23400	3.00000
	S.D.	0.81777	0.92534	0.32601	0.61237
E_2	$\bar{x}$	6.00000	5.15000	3.90000	3.50000
	S.D.	0.35355	1.11243	1.08832	1.22474
E_3	$\bar{x}$	5.55000	4.45000	2.73400	2.20000
	S.D.	0.97468	0.83666	0.25255	0.57009
E_4	$\bar{x}$	5.00000	4.58250	3.33250	2.37500
	S.D.	1.47196	1.32309	2.32411	1.88746

Table 2.45

Analysis of Variance for Experimental Groups' Ratings of Disruptiveness

Source	SS	DF	MS	F
Mean	1218.21948	1	1218.21948	360.40698
Group (A)	10.18668	3	3.39556	1.00457
Error	50.70183	15	3.38012	
Phase (B)	83.84337	3	27.94779	61.92337***
A x B	3.97388	9	0.44154	0.97832
Error	20.30978	45	0.45133	

*** $p < .001$

manner. In any event, it is clear that this relatively global measure of perceived behavior change showed that target children greatly reduced their disruptiveness over time. In order to determine whether this effect was largely attributable to the intervention, the global ratings of untreated control subjects were examined. The five acting-out children in group C_1 received a mean rating of 5.2 during baseline and 4.6 at the conclusion of the study (see Table 2.46). Results of the second ANOVA in which initial and final ratings of all five groups were compared are presented in Table 2.47. A significant main effect was obtained for occasions (F = 87.197, df = 1, 19; $p < .001$) but not for groups (F = 1.309, df = 4, 19; n.s.). A significant interaction was obtained (F = 3.906, df = 4, 19; $p < .02$). Orthogonal comparisons between post-treatment mean ratings showed that groups E_1-E_4 were perceived as significantly less disruptive than their untreated counterpart, C_1 (F = 16.575, df = 1, 19; $p < .01$).

Distractibility. Table 2.48 gives the mean ratings and standard deviations by phase for the four experimental groups. During baseline, teachers' summary reports showed an average rating of 5.43 on a seven-point scale. This decreased to 4.37 during early intervention (i.e., "moderately distractible"), 3.68 in later intervention, and 3.14 at follow-up. The ANOVA (see Table 2.49) yielded significant main effects for group (F = 4.025, df = 3, 15; $p < .03$) and phase (F = 24.542, df = 3, 45; $p < .001$). Group differences were largely attributable to non-equivalent baselines. In no case does this difference suggest the existence of a tracking effect. It seems very clear that distractibility, as measured in a relatively global fashion, decreased substantially over time. Ascribing such improvement to the intervention procedures necessitates the examination of control group data. Summary reports submitted by

Table 2.46

Initial and Final Ratings for Summary Reports of Disruptiveness

Group		Initial rating	Final rating
E_1	$\bar{x}$	5.80000	3.00000
	S.D.	1.30384	0.70711
E_2	$\bar{x}$	6.00000	3.40000
	S.D.	0.0	1.34164
E_3	$\bar{x}$	5.80000	2.00000
	S.D.	1.30384	0.70711
E_4	$\bar{x}$	5.25000	2.25000
	S.D.	1.70782	1.89297
C_1	$\bar{x}$	5.20000	4.60000
	S.D.	0.83666	0.89443

Table 2.47

Analysis of Variance for Initial and Final Summary Reports of Disruptiveness

Source	SS	DF	MS	F
Mean	892.80176	1	892.80176	509.40747
Group (A)	9.17914	4	2.29478	1.30934
Error	33.29993	19	1.75263	
Occasion (B)	78.01843	1	78.01843	87.19746***
A x B	13.97905	4	3.49476	3.90593*
Error	16.99992	19	0.89473	

* $p < .05$

*** $p < .001$

Table 2.48

Experimental Group Ratings of Distractibility

Group		Phases I and III Baseline	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	$\bar{x}$	5.30000	3.86600	3.53400	3.25000
	S.D.	0.77862	0.80764	0.76862	1.03078
E_2	$\bar{x}$	6.15000	5.31600	4.60000	4.10000
	S.D.	0.37914	1.10929	1.23468	1.38744
E_3	$\bar{x}$	4.50000	3.65000	2.56600	2.00000
	S.D.	1.33463	1.49583	0.43506	0.35355
E_4	$\bar{x}$	5.87500	4.70750	4.08500	3.25000
	S.D.	0.52042	1.03789	2.04360	2.25462

Table 2.49

Analysis of Variance for Experimental Groups' Ratings of Distractibility

Source	SS	DF	MS	F
Mean	1310.43823	1	1310.43823	427.49243
Group (A)	37.01360	3	12.33786	4.02487*
Error	45.98111	15	3.06541	
Phase (B)	55.81955	3	18.60651	24.54240***
A x B	1.98495	9	0.22055	0.29091
Error	34.11617	45	0.75814	

* $p < .05$

*** $p < .001$

by teachers in group C_1 showed a mean of 6.20 during the baseline period and 5.80 at the termination of the study (see Table 2.50). The latter figure was considerably higher than the mean final rating given experimental subjects (3.00). The ANOVA findings are presented in Table 2.51. Orthogonal comparisons showed this difference to be significant (F = 19.169, df = 1, 19; p < .01). Therefore, untrained teachers continued to rate their children as quite distractible while trained teachers considered their target students to be functioning at better than a "moderate" distractibility level.

Social withdrawal. Findings for the subsample of socially withdrawn children (N = 25) were somewhat similar to results obtained for disruptive/distractible students. Table 2.52 shows that teacher ratings during baseline were consistently high for experimental groups. The baseline mean across these four groups was 5.44. A decrease to 4.90 coincided with early treatment and during late intervention the mean level fell to 3.53. Considerable continued improvement was observed during the brief follow-up period, resulting in an average rating of only 2.44. The ANOVA for repeated measures (see Table 2.53) yielded a significant main effect for phase of treatment (F = 94.549, df = 3, 48; p < .001). Target children in experimental groups were, therefore, perceived as improving throughout the course of the study. The corresponding experimental and control group means from the first weekly rating were 5.5 and 5.0 respectively. At termination the children whose teachers received training were rated as less withdrawn ($\bar{x}$ = 2.1) than target children in group C_1 ($\bar{x}$ = 5.4). Members of the latter group actually appeared to be somewhat more withdrawn than in baseline (see Table 2.54). The ANOVA (see Table 2.55) produced significant main effects for both group (F = 5.103, df = 4, 20; p < .01) and occasion (F = 207.429, df = 1, 20; p < .001). The interaction

Table 2.50

Initial and Final Ratings for Summary Reports of Distractibility

Group		Initial rating	Final rating
E_1	$\bar{x}$	5.20000	3.20000
	S.D.	1.30384	0.83666
E_2	$\bar{x}$	6.20000	3.80000
	S.D.	0.83666	1.30384
E_3	$\bar{x}$	4.40000	2.20000
	S.D.	1.51657	0.44721
E_4	$\bar{x}$	6.50000	2.75000
	S.D.	1.00000	2.21736
C_1	$\bar{x}$	6.20000	5.80000
	S.D.	0.83666	1.09544

Table 2.51

Analysis of Variance for Initial and Final Summary Reports of Distractibility

Source	SS	DF	MS	F
Mean	1018.59790	1	1018.59790	820.92969
Group (A)	38.67505	4	9.91876	7.99394**
Error	32.57494	19	1.24079	
Occasion (B)	55.02942	1	55.02942	33.75502***
A x B	12.94157	4	3.23539	1.98459
Error	30.97493	19	1.63026	

** $p < .01$

*** $p < .001$

Table 2.52

Experimental Group Ratings of Withdrawal

Group		Phases I and III Baseline	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	$\bar{x}$	6.10000	5.65000	4.10000	3.10000
	S.D.	0.76240	0.96177	1.53279	1.55724
E_2	$\bar{x}$	5.40000	4.60000	3.40000	2.45000
	S.D.	1.25748	2.06610	2.09868	1.62404
E_3	$\bar{x}$	4.50000	4.20000	2.93400	1.70000
	S.D.	0.50000	0.37081	0.93206	1.09544
E_4	$\bar{x}$	5.75000	5.15000	3.70200	2.50000
	S.D.	0.35355	0.74162	0.60677	0.61237

Table 2.53

Analysis of Variance for Experimental Groups' Ratings of Withdrawal

Source	SS	DF	MS	F
Mean	1329.91553	1	1329.91553	289.36377
Group (A)	20.83112	3	6.94370	1.51081
Error	73.53598	16	4.59600	
Phase (B)	110.22121	3	36.74040	94.54903***
A x B	0.84512	9	0.09390	0.24165
Error	18.65211	48	0.38859	

*** $p < .001$

Table 2.54

Initial and Final Ratings for Summary Reports of Withdrawal

Group		Initial rating	Final rating
E_1	$\bar{x}$	6.00000	2.40000
	S.D.	1.00000	1.51657
E_2	$\bar{x}$	5.60000	2.60000
	S.D.	1.14017	1.51657
E_3	$\bar{x}$	4.60000	1.20000
	S.D.	0.54772	0.44721
E_4	$\bar{x}$	5.80000	2.20000
	S.D.	0.44721	0.44721
C_1	$\bar{x}$	5.00000	5.40000
	S.D.	1.11803	0.54772

Table 2.55

Analysis of Variance for Initial and Final Summary Reports of Withdrawal

Source	SS	DF	MS	F
Group (A)	26.6800	4	6.67000	5.13077**
Error	26.0000	20	1.30000	
Occasion (B)	87.1200	1	87.12000	207.42900***
A x B	29.4800	4	7.37000	17.54760***
Error	8.40000	20	0.42000	

** $p < .01$

*** $p < .001$

term was also significant ($F = 17.548$, $df = 4, 20$; $p < .001$. When the termination means for the four experimental groups were contasted with the final rating for C_1, a highly significant difference was obtained ($F = 103.714$, $df = 1, 20$; $p < .01$). Once again, no pattern of results was obtained which would be predicted on the basis of differential amounts of observation training, practice, or monitoring (Group E_1, which received no such input, compared favorably on all measures).

To summarize, the summary report measure showed substantial improvement for experimental target children on all three dependent variables. Untreated control children showed slight (non-significant) improvement from the onset to the end of the study on dimensions of disruptiveness and distractibility. Socially withdrawn children in Group C_1 were actually perceived as more handicapped at termination than they were five months earlier.

Behavior Vignettes Test

Group means and standard deviations are presented in Table 2.56. Pre-test scores across the four experimental groups had a mean of 7.77 correct out of the 20 items. Teachers in group C_1 scored about the same ($\bar{x}$ = 7.40). At termination, the mean for trained teachers was 11.72, a 33% improvement over baseline. The control group scored slightly lower in the post-test than they had earlier ($\bar{x}$ = 6.70).

A 5 (group) x 2 (occasion) ANOVA for repeated measures was performed on the number of items correct. Unlike previously discussed factorial designs, the subsamples of teachers whose target child was either socially withdrawn or disruptive/distractible were combined. Theoretically, knowledge of behavioral principles and strategies should not have been directly affected by either the type of target child or the fact that one target child dropped out. Consequently, each of the 50 teachers provided data on both occasions.

Significant main effects were obtained for both groups (F = 3.189, df = 4, 45; $p < .02$) and occasions (F = 52.150, df = 1, 45; $p < .001$). Table 2.57 also shows that the interaction term attained significance (F = 5.474, df = 4, 45; $p < .001$). Orthogonal comparisons between pre- and post-treatment scores for the four experimental groups showed a significant gain (F = 71.371, df = 1, 45; $p < .01$). When post-test means for E_1-E_4 were contrasted with that of C_1, a significant difference was also demonstrated (F = 46.202, df = 1, 45; $p < .01$). Therefore, teachers who received behavior modification training increased their knowledge of basic principles and techniques to the point where they were clearly more advanced than teachers who did not receive training.

Table 2.56

Group Means and Standard Deviations on Behavior Vignettes Test

Group		Pre	Post
E_1	$\bar{x}$	7.50000	11.00000
	S.D.	4.79003	3.05505
E_2	$\bar{x}$	7.70000	12.80000
	S.D.	3.33499	3.29309
E_3	$\bar{x}$	7.60000	11.00000
	S.D.	2.63312	1.66666
E_4	$\bar{x}$	8.30000	12.10000
	S.D.	3.12872	2.85385
C_1	$\bar{x}$	7.40000	6.70000
	S.D.	2.11870	1.41813

Table 2.57

Analysis of Variance for Behavior Vignettes Test Scores

Source	SS	DF	MS	F
Group (A)	134.740	4	33.6850	3.18886*
Error	475.350	45	10.5633	
Occasion (B)	228.010	1	228.0100	52.14970***
A x B	95.740	4	23.9350	5.47433**
Error	196.750	45	4.3722	

* $p < .05$

** $p < .01$

*** $p < .001$

Number of Programs Implemented

A total of 162 operant programs supplemented the 39 which were specifically designed and implemented for target children. During the tenure of the study, teachers carried out an average of 5.03 programs (s.d. = 3.35). Only two teachers reported that they chose not to extend their systematic practice of behavior modification beyond application to the target child.

Programs were characterized along two dimensions: (a) appropriate for use either with an individual student or group, and (b) having contingencies attached to either behaviors correlated with academic performance (e.g., volunteering, noncompliance) or to the performance itself (e.g., number of assignments completed, percentage of problems correct). Table 2.58 gives the frequencies for each type of supplemental program.

Table 2.58

Frequencies of Additional Programs

	Individual	Group	
Behavioral	76	44	120
Academic	33	9	42
	109	53	N = 162

Clearly, there was a preference for dealing with conduct problems at an individual level. Such a tendency might well be attributed to the fact that the initial program (for the target child) was directly focusing on behaviors rather than academic performance. Hence, teachers may have become more

comfortable in administering programs of this type. Had the sample included teachers of special education, educable retarded, or learning disabilities classes, one might expect greater emphasis on programs utilizing group contingencies.

Of interest are the large group differences in terms of introduction of additional programs. Table 2.59 shows that teachers in group E_2 introduced an average of 6.6 additional programs compared to 4.6 for E_4, 3.0 for E_3, and only 2.0 for E_1.

Table 2.59

Number of Additional Programs Implemented in Each Experimental Group

Group	$\bar{x}$	S.D.
E_1	2.00	1.25
E_2	6.60	4.09
E_3	3.00	1.94
E_4	4.60	3.41

To the extent that such self-report data can be trusted, the operant training program appears to have had an effect on a large number of children experiencing a wide range of difficulties.

Global Rating

At follow-up, 35 of the 39 experimental teachers (90%) rated their target child as having improved "a great deal." Each of the groups (E_1-E_4) had one teacher who rated her target child as "somewhat improved."

Expectations of Improvement

Table 2.60 gives the experimental group means and standard deviations for teacher expectations of improvement. Probabilities during baseline averaged about 70%. Expectations rose to a high level during early intervention (81%), and continued to increase through the later stages of treatment (84%). A slight decline to 79% was observed during follow-up, possibly due to the short period remaining in the school year or the fact that many children had already improved to the point where they were indistinguishable from their peers. A 4 (group) x 4 (phase) ANOVA for repeated measures was conducted on these probability estimates (see Table 2.61). A significant main effect was obtained for phase ($F = 30.379$, $df = 3, 105$; $p < .001$) indicating that anticipated success increased as training progressed. The generally low standard deviations reflect a consensus among teachers that the daily problems they were facing would be ameliorated as a function of their participation in the program.

A breakdown of the probabilities by group shows that E_1 was the least optimistic throughout the course of the study. This was reflected in a significant main effect for groups ($F = 2.819$, $df = 3, 35$; $p < .05$). One feasible explanation for E_1's lower estimate during baseline is that the placebo input which it received in lieu of observation training was perceived as irrelevant to the task at hand. E_1 did show a 12% increase from baseline to follow-up, a change commensurate with that of the other groups. However, its higher standard deviation suggests that a ceiling effect was operating on the other groups in such a way as to invalidate the instrument as one with ratio scale properties. In other words, the difference between a 10% increase from 50%

Table 2.60

Percent Probability of Expected Improvement

Group		Phases I and III Baseline	Phase IV Early intervention	Phase V Late intervention	Phase VI Follow-up
E_1	$\bar{x}$	64.50000	67.33997	73.65993	76.00000
	S.D.	16.65833	16.91417	17.23563	18.37872
E_2	$\bar{x}$	72.66998	81.49994	84.65993	86.00000
	S.D.	9.65862	11.62402	8.88446	9.66091
E_3	$\bar{x}$	78.00000	82.00998	86.32996	88.00000
	S.D.	11.35292	7.88999	6.37688	5.86894
E_4	$\bar{x}$	70.83333	76.31108	84.81105	86.11110
	S.D.	7.60345	8.07069	10.68952	16.72903

Table 2.61

Analysis of Variance for Expected Improvement

Source	SS	DF	MS	F
Mean	963491.62500	1	963491:62500	2036.06787
Group (A)	4001.88672	3	1333.96216	2.81895*
Error	16562.42188	35	473.21191	
Phase (B)	3786.43750	3	1262.14575	30.37949***
A x B	219.14844	9	24.34992	0.58609
Error	4362.32813	105	41.54597	

* $p < .05$

*** $p < .001$

to 60% is not equivalent to a change from 80% to 90%. Hence, analysis of gain scores would not be appropriate.

The result that members of E_1 were somewhat more skeptical would be troublesome were it not for the fact that this group performed competitively on virtually all other dependent measures. Only on the number of additional programs were they lacking. On the basis of actual and perceived behavior change of target children, E_1 showed no deficiencies whatever. It would seem that the level of expectation for E_1, though lowest, was sufficient to permit cooperative, effective implementation of prescribed procedures.

Cost Analysis

The cost of implementing a teacher training program identical to that used here, but without research components (e.g., professional observer's salaries and expenses, control group, data analysis) was calculated. Table 2.62 gives

estimates for consulting psychologist's time and estimated expenditures, with the hourly rate set at $6.75 per hour.[19] Only 8.4 hours of professional time per subject were required (at a cost of $65.44). Training additional groups during the same period of time obviates the need for further preparation and, hence, reduces the time allotment to 7.5 hours per subject (at a cost of $59.36). Estimates of teacher time and expenditures are given in Table 2.63. Thirty-four and one-half hours of out-of-class work were required (at a cost of $23.70). Were training conducted through a university, an additional fee could be assessed per hour of credit.

Table 2.62

Cost of Conducting In-Service Teacher Training

Item	Psychologist time per teacher	Cost[19]
Observations (2)	1.5 hr.	$ 10.13
Contact hours (20 per teacher; 10 teachers per group)	2.0	13.50
Preparation for group session[20]	.9	6.08
Phone consultation	1.5	10.13
Review teacher-collected data	.5	3.38
Transportation[21]	2.0	18.72
Supplies, reprints, postage	-	3.50
Total per subject	8.4 hr.	$ 65.44

[19] Estimates based on 1974 Quebec salary schedule for school psychologist (M.A.) with five years' experience; annual salary $13,000 ($6.75/hr.).

[20] Estimate based on training one group only. Additional groups would not require this item.

[21] Based on 2.9 round trips of 15 miles each and 40 minutes (12¢ per mile).

Table 2.63

Cost to Participants in Program

Item	Teacher time	Cost
Contact hours (10 two-hour sessions)	20.0 hr.	-
Preparation for session	9.0	-
Phone consultation	1.5	-
Transportation	6.0	$ 16.20
Textbook	-	7.50
	34.5	$ 23.70[22]

[22] As of 1974, estimated cost of training through McGill University Department of Continuing Education would include an additional $60.00 for three credits.

Discussion

The present thesis represents one of the few attempts to simultaneously test the same hypothesis in both the laboratory and naturalistic settings. While it is ofen espoused that generalizing from analogue research is a tenuous proposition, empirical support for this position is generally lacking (Orne, 1962).

Results of Experiment I confirmed findings from earlier laboratory studies in which teachers trained in and practicing systematic observation techniques submitted ratings of child behavior which corresponded with observed operant levels. Individuals who received no such training made ratings of poor convergent validity.

Experiment II provided a naturalistic test for this "tracking effect" on perception of child behavior change. Unlike results obtained in the laboratory, existence of such a phenomenon was not evident in the field. Teacher ratings of distractibility, disruptiveness, and social withdrawal failed to converge with independently observed behavioral levels as a function of observation training, daily data collection, or external monitoring. Therefore, Hypothesis I was rejected.

A number of factors operating either in isolation or together may have accounted for this result. The teacher training program may have produced an expectancy for an emerging pattern of child behavior. Such an influence was not present in the early stages of the analogue study;[23] however, it may have

---

[23] Evidence of a "halo effect" in three of the four groups in Experiment I suggests that an expectation of improvement had been established by the sixth trial.

been of sufficient impact to override the effects of systematic observations in the actual classroom. The data which teachers recorded may, in fact, have been reliable; judging from comparisons of data collected by teachers in E_4 and their external monitors, this was the case. However, these data may not have played a major role in formulating the teachers' overall appraisal of a condition such as social withdrawal. The tendency for individuals to ignore behavior frequency data in generating a global appraisal has been demonstrated (Hines, 1974; Walter & Gilmore, 1973).

Another strong possibility is that group differences were not obtained because teachers in each group were, in fact, tracking during the intervention phase. Virtually all behavior modification techniques require that individuals pinpoint and observe certain target behaviors and react to these in a systematic, predetermined fashion. Hence, the treatment programs themselves contain an element of "self-training" in observation. Certainly, the use of behavior modification by all four experimental groups served to reduce any variance which may have emerged due to situational learning of observation skills. It seems reasonable to assume that differences would be magnified in cases where behaviorally oriented treatment is compared to more conventional counseling or phenomenological approaches. However, because these approaches have not generally been as successful in dealing with classroom management problems (O'Leary & O'Leary, 1972), it would have been difficult to justify their use merely for purposes of testing for a tracking effect.

Some indirect support for the notion that behavior modification programs reduce individual differences in observation skills comes from the untreated control group. Withdrawn target children in C_1 did show improvement on observation data but were not rated as such either on the Walker Checklist or

summary reports. In fact, the summary report indicated slight deterioration. Conversely, acting-out children were rated as improved on both perceptual measures, but did not improve appreciably according to direct observations. Therefore, for both subsamples of control children, there were discrepancies between measures. For experimental subjects, results were consistent across all criterion variables. This consistency might well be attributed to the fact that all experimental groups were indeed tracking during the intervention phase.

Equivalent findings for experimental groups were not confined to teacher perception. For acting-out children no group differences were found either in target child behavior or in teacher attention. Group E_1, which was not trained in observation techniques, performed very well on all measures. Some evidence of a tracking effect was found in analyzing data for the socially withdrawn subsample. Following intervention, teachers who were not collecting data on a daily basis reduced substantially their attention to target children. Concomitantly, a sharp increase occurred in self-stimulation of target children in groups E_1 and E_2. Therefore, when a withdrawn child displays high base rates of self-stimulation or looking around, it could be recommended that his teachers collect data regularly (at least once or twice per week). This is the only situation for which there exists empirical evidence that data collection may be advantageous. Acceptance of this finding would, no doubt, be considerably enhanced by two factors: (1) a longer follow-up period (although by and large, teachers, parents, or other trainees are seldom asked to collect additional data following the termination of treatment), and (2) empirical evidence showing a functional relationship between teacher attention and self-stimulation.

Despite the inclusion of data collection in virtually all behavior modification teacher training programs, there seems to be no reason to require such

effort when treatment is focused on disruptive-distractible children or socially withdrawn children whose only deficiency lies in the area of peer interaction.

The suggestion that observation training and data gathering by teachers be eliminated in most cases is not necessarily generalizable to the family situation. Because teachers have a greater normative base against which to evaluate child behavior, it is quite possible that their perceptions converge with actual behavior levels to a greater degree than do the impressions of parents. Consequently, any alterations made on the basis of these results should be confined to the school environment.

Demand Baseline Procedure

Hypothesis II stated that child behavior would improve over baseline levels under conditions of high teacher motivation. This hypothesis was partially confirmed. Socially withdrawn children emitted considerably more prosocial and attending behaviors during the high demand condition than during natural baseline. This shift was accompanied by large increases in both absolute and relative teacher attention directed toward the target child. Whether teachers could continue this pattern on a regular basis is difficult to assess. Certainly, one cannot rule out the possibility.

Omitting the demand condition would have resulted in overestimating the degree of impact ascribed to the behavior modification techniques themselves. Because social withdrawal at follow-up was significantly lower than observed in the demand phase, one can conclude that just raising teacher motivation would not alone produce improvement of the magnitude generated by the teacher training. However, a less powerful intervention might well be perceived as effective given

a concomitant rise in demand characteristics. For this reason, it is recommended that some of the following precautions be taken in future research: (a) a high demand procedure be introduced in future studies of socially withdrawn children, (b) a correction factor be applied to remove the variance attributed to differential demand characteristics which may operate as a function of treatment phase, or (c) an attention placebo or pseudotherapy control group be included in order to assess the effects of high expectations.

The total inability of teachers to decrease disruptive or distractible behavior in target children during the demand phase was not predicted. This finding is consistent with that obtained by Johnson and Lobitz (1974), who found that parents of deviant children could not make their children "look good" when instructed to do so. Notwithstanding, it was presumed that classroom teachers were generally more skillful in this regard than parents in deviant families. The fact that they were not suggests that teachers may be presenting similar facilitating stimuli for noxious behavior. Moreover, it may be that acting-out children shape their teachers into attending to deviant behavior in much the same way they do their parents. In any event, results of the demand baseline procedure do not recommend manipulation of demand characteristics in order to isolate the effects of treatment techniques designed for children labeled "acting out."

Evaluation of Intervention: Acting-Out Children

Results from multiple measures were consistent in showing that the present intervention was efficient, inexpensive, and effective in dealing with disruptiveness and distractibility.

Table 2.64 shows the results of analysis of observation data, the WPBIC,

Table 2.64

Summary of Results for Disruptiveness

Dependent variable/ comparison	Observation data (PA)+(DI)+(T-)+(P-)	Walker Checklist Acting-out scale	Summary report Disruptiveness	Hypothesis as stated:
Intra-group: Baseline vs. Follow-up Hypothesis III	Experimental groups (E_1-E_4) improved significantly ($p < .01$) The control group (C_1) did not	Both (E_1-E_4) and C_1 improved significantly ($p < .05$)	E_1-E_4 improved significantly ($p < .01$) C_1 did not	Accepted
Inter-group: Experimentals (E_1-E_4) vs. Control (C_1) at Follow-up Hypothesis IV	E_1-E_4 were significantly less disruptive than C_1 ($p < .01$)	E_1-E_4 and C_1 did not differ significantly	E_1-E_4 were significantly less disruptive than C_1 ($p < .01$)	Accepted
Normative: Target children vs. Peers/norm at Follow-up Hypothesis V	Target children in E_1-E_4 did not differ from peers Target children in C_1 were significantly more disruptive than their peers ($p < .05$)	E_1-E_4 and C_1 scored within one standard deviation of the norm	N/A	Accepted

and weekly summary reports for disruptiveness. One can readily see that experimental subjects fared well on seven out of eight measures compared to only three of eight for C_1. Both experimental and control groups showed lower ratings which were within the normal range on the second administration of the WBPIC. All comparisons on the other two measures differentiated experimental groups from their control. Reading across the rows of Table 2.64 shows that each of the three relevant hypotheses (see pp. 49-50) were accepted, based on confirmation by the observation data and at least one of the other two measures.

Results for distractibility are even more impressive. Table 2.65 shows that experimental target children improved on seven of eight criterion variables compared to only two for the students in group C_1. Again, untreated children were reported as improved from pre-test to post-test on the WPBIC, and were statistically indistinguishable from treated children at follow-up. However, members of C_1 were still one standard deviation more distractible than the normative sample.

To summarize, acting-out children whose teachers received training improved significantly from baseline to termination. At follow-up, they also differed significantly from matched control students whose teachers did not actively participate and were indistinguishable from their normal peers. It is of interest that children in C_1 were rated on the WPBIC as improved despite the fact that direct observation and the teachers' own global impressions showed no change over time. Such a finding indicates a possible practice effect for this instrument, a deficiency not uncommon to teacher rating scales (Sprague, Christiansen, & Werry, 1972; Spivack & Swift, 1973). Test-retest reliability was not reported in the WPBIC Manual (Walker, 1970). Perhaps the instrument should be restricted to use as a diagnostic tool vis-a-vis

Table 2.65

Summary of Results for Distractibility

Dependent variable/ comparison	Observation data (HR)+(P-)+(IL)+(SS)+ (LO)+(NA)	Walker Checklist Distractibility scale	Summary report Distractibility	Hypothesis as stated:
Intra-group: Baseline vs. Follow-up Hypothesis III	Experimental groups (E_1-E_4 improved significantly ($p < .01$) The control group (C_1) did not	Both (E_1-E_4) and C_1 improved significantly ($p < .01$)	E_1-E_4 improved significantly ($p < .01$) C_1 did not	Accepted
Inter-group: Experimentals (E_1-E_4) vs. Control (C_1) at Follow-up Hypothesis IV	E_1-E_4 were significantly less disruptive than C_1 ($p < .01$)	E_1-E_4 did not differ significantly from C_1	E_1-E_4 were significantly less distractible than C_1 ($p < .01$)	Accepted
Normative: Target children vs. Peers/norm at Follow-up Hypothesis V	Target children in E_1-E_4 did not differ from peers Target children in C_1 were significantly more distractible than their peers ($p < .01$)	E_1-E_4 scored within one standard deviation above the norm	N/A	Accepted

a criterion measure. While Walker (1970) made no claims as to its value as a dependent variable, factor scores have been used in this manner (Walker, Hops, Greenwood, & Todd, 1975).

The intervention did not require that teachers devote an inordinate amount of time to the target child. These children typically garner more than their fair share of attention; however, it was not necessary for teachers to increase the relative or absolute quantity of attention for the purpose of inducing behavior change. While this is not an original finding, it does address one of the most frequently raised objections to teacher-mediated treatment. Generally, teachers express concern about ability to carry out a program which ostensibly requires increased contact with a particular child. Using the same amount of attention contingently is usually all that is necessary. That teachers simultaneously introduced a large number of additional programs is strong evidence that such doubts were allayed.

Table 2.66 summarizes results using the composite withdrawal score, the withdrawal factor score of the WPBIC, and the seven-point summary report. The intra-group comparisons are consistent across measures in showing that target children whose teachers were trained as behavior modifiers improved significantly from baseline to follow-up. Hypothesis VI was therefore accepted. Students in group C_1 showed non-significant improvement over time on the observational measure. Yet, this was of sufficient magnitude to render C_1 statistically indistinguishable from experimental groups at follow-up. Hence, Hypothesis VII, which predicted differences between treated and untreated groups, was rejected, despite the fact that significant differences were obtained on the WPBIC and summary reports. With regard to Hypothesis VIII, neither treated nor untreated target children attained peer levels of prosocial

Table 2.66

Summary of Results for Social Withdrawal

Dependent variable/ comparison	Observation data (P+) (VO)+(IT)- (SS)-(LO)	Walker Checklist Withdrawal scale	Summary report Withdrawal	Hypothesis as stated:
Intra-group: Baseline vs. Follow-up Hypothesis VI	E_1-E_4 improved significantly over natural baseline ($p < .01$) and demand baseline ($p < .02$) C_1 did not improve significantly over baseline	E_1-E_2 improved significantly ($p < .01$) C_1 did not	E_1-E_4 improved significantly ($p < .01$) C_1 did not	Accepted
Inter-group: Experimentals (E_1-E_4) vs. Control (C_1) at Follow-up Hypothesis VII	E_1-E_4 were not significantly less withdrawn than C_1	E_1-E_4 were significantly less withdrawn than C_1 ($p < .01$)	E_1-E_4 were significantly less withdrawn than C_1 ($p < .01$)	Rejected
Normative: Target children vs. Peers/norms at Follow-up Hypothesis VIII	Target children in E_1-E_4 were significantly more withdrawn than peers ($p < .01$) Target children in C_1 were also more withdrawn than peers ($p < .05$)	E_1-E_4 scored within one standard deviation of the norm C_1 did not	N/A	Rejected

and attending behavior. Still, three of the four experimental groups did not differ from their peers despite the overall (across group) differences. The WPBIC showed students in E_1-E_4 were perceived as no longer withdrawn, while children in C_1 still scored one standard deviation above the norm. Since the criterion for acceptance includes the support of observation data, Hypothesis VIII was rejected.

Not only do the three measures differ on dimensions of retrospection, specificity, and independence of raters (observers), but also on that of content. The withdrawal composite consists of five discrete prosocial or off-task behaviors of unit weight (see Appendix B). The WPBIC withdrawal scale focuses exclusively on peer interaction. It includes no items related to volunteering, initiation to teacher, self-stimulation, or looking around. Yet, such behaviors do discriminate withdrawn children from their peers (Bell, Waldrop, & Weller, 1972; Camp & Zimet, 1974). In fact, 43% of the withdrawn children treated by Walker and Hops from 1973-1975 have shown excessive daydreaming (personal communication).[24] In any event, scores on the WPBIC withdrawal scale should correlate higher with a behavioral measure of interaction with peers than with other responses or a composite of categories. The summary report may also have disproportionately weighted the interaction component since many teachers selected this area as being of most concern to them. Certainly, the

---

[24] This finding is based on having the following item checked by teachers completing the WPBIC: "Frequently stares blankly into space and is unaware of his surroundings when doing so." This item loaded largely on the distractibility factor when the original validity studies were conducted. On the basis of later findings by both Hops and Walker and this author, there is good reason to include this item in the Withdrawal Scale.

thrust of teacher training was directed to the problem of inadequate peer relationships.

When the observational dependent variable was reduced to "appropriate interaction with peers" (P+), the results of the program become much more consistent across measures. Table 2.67 shows that all eight comparisons significantly favor the children in experimental groups over their untreated counterparts. The intervention was successful in raising the level of peer interaction to that maintained by peers. Withdrawn children whose teachers received no training did not significantly improve their level of interaction on any of the three variables. Hypotheses VI, VII, and VIII were accepted, but only when restricted to their peer interaction component.

Results from analyses of observation data showed that volunteering and initiations to teacher did not increase for experimental groups. This may have been accounted for, in part, by the relatively poor academic performance of the withdrawn children. While generally not failing, they were consistently rated as below average in most subjects. Few children, withdrawn or otherwise, will raise their hand when they do not know the answer. Initiating to teacher may not have changed because target children in E_1-E_4 were explicitly taught and encouraged to approach their peers when assistance was needed.

Observational data also show substantial reductions in self-stimulation and looking around in both the experimental and control groups. Hence, it appears that the normal educational process and/or maturational factors are likely to produce increased on-task behavior in 7-9-year-old, socially withdrawn children.

Withdrawn children in group C_1 were actually rated by teachers as more handicapped at termination than at baseline on both the WPBIC and summary

Table 2.67

Summary of Results for Appropriate Interaction with Peers

Dependent variable/ comparison	Observation data P+	Walker Checklist Withdrawal scale	Summary report Withdrawal	Hypothesis as stated:
Intra-group: Baseline vs. Follow-up - Hypothesis VI	E_1-E_4 improved significantly over natural baseline ($p < .01$) and demand baseline ($p < .01$) C_1 did not improve significantly over baseline	E_1-E_4 improved significantly ($p < .01$) C_1 did not	E_1-E_4 improved significantly ($p < .01$) C_1 did not	Accepted
Inter-group: Experimentals (E_1-E_4) vs. Control (C_1) at Follow-up Hypothesis VII	E_1-E_4 interacted significantly more than C_1 ($p < .01$)	E_1-E_4 were significantly less withdrawn than C_1 ($p < .01$)	E_1-E_4 were significantly less withdrawn than C_1 ($p < .01$)	Accepted
Normative Target children vs. Peers/norms at Follow-up Hypothesis VIII	Target children in E_1-E_4 did not differ significantly from peers Target children in C_1 did not differ significantly from peers ($p < .15$)	E_1-E_4 scored within one standard deviation of the norm C_1 did not	N/A	Accepted

reports. Conversely, the behavioral data indicated slight improvement in peer interaction as mentioned above, and substantial decreases in self-stimulation and looking around. Such a discrepancy between measures, while not large in magnitude, lends support to the notion that withdrawn children are more or less ignored by their environment. Once a teacher forms an impression, she seems rather resistant to change.

An examination of phase means of each of the observational composites shows that changes in early intervention were more commonly associated with acting-out than socially withdrawn children. The former group improved more rapidly. There are several possible explanations for this: (a) contingencies for deviant behavior were applied earlier in the program; teachers of withdrawn children followed a lengthy sequence which included symbolic modeling, role-playing, and self-monitoring prior to the introduction of specific contingencies; (b) acting-out children were not usually required to learn new skills but to increase behaviors already existing in their repertoires and to reduce their rates of noxious responses. A number of socially withdrawn children were clearly deficient in social skills and required gradual shaping of new prosocial behaviors (e.g., initiating to a group of peers); (c) the incentive for teachers to control disruptive behavior may well be greater than for shaping social approach skills. Acting-out children often emit behaviors which "demand" or "force" an environmental response (e.g., consequence), whereas withdrawn youngsters may display a positive behavior which goes unnoticed. Hence, the acting-out child is more likely to receive consistent, intensive feedback from the teacher. Teachers and counselors should be aware that the probability of a socially withdrawn child improving immediately is not as high as one would expect of a coercive child.

In summary, the intervention was most effective in increasing peer interaction in socially withdrawn children who, if left untreated, would probably not have improved appreciably. The long-term stability of this condition has been demonstrated by Waldrop and Halverson (1975), who found that preschool children who have restricted peer relations exhibit the same pattern five years later. It was also demonstrated that preschool children who were not assertive at first testing in dealing with an experimental barrier situation, also had difficulty five years later in coping with novelty and were ill at ease with peers (Halverson & Woldrop, 1974). It seems that a child "gets back what he puts out" (Kohn, 1966), that he evokes from peers the kind of responses that will enable him to maintain his prevailing routine and that this style of life will continue into adulthood (Michael, Morris, & Soroker, 1957; Morris, Soroker, & Buruss, 1954). This is not to imply that socially withdrawn children are likely to become seriously disturbed adults. The paucity of follow-up data suggests that withdrawn individuals who had received brief counseling at a public mental health clinic were reasonably well-adjusted in adulthood (an average of 26 years later). They tended to lead quiet, retiring lives characterized by stability both at work and in marriage. Social contacts were restricted but certainly not rare. The incidence of psychotic disorders was no greater than one or two percent in the two studies cited. There is no evidence to indicate that untreated withdrawn children would fare any worse. In fact, the authors attribute little improvement to the brief treatment, suggesting instead that the clinical concern with social withdrawal has been exaggerated. Still, it is apparent that withdrawn individuals are handicapped in that they avoid many potential opportunities for social learning. While there are other disorders of higher

priority, an efficient, low-cost treatment procedure would seem well worth implementing.

The present intervention increased not only the proportion of time target children interacted, but also their rate of initiation to peers, the range of contacts and the average duration of conversations. The last three variables were evaluated solely on the basis of teacher-collected data (groups E_3 and E_4), which, albeit reliable (relative to the external monitor) may be biased. There is also anecdotal evidence which suggests that these changes were not restricted to the classroom, but generalized to the playground and lunchroom. It was reported that target children began to invite friends home after school and were invited more often themselves. These children were described as happier, more alert, and were said to be bringing in toys to show others, smiling and laughing more, and assuming tasks of greater responsibility.

Because this study suffers from a brief follow-up, it is impossible to draw any definitive conclusions about generalization of effects over time. However, Walker, Hops, Greenwood, and Todd (1975) report excellent six-month follow-up after using very similar treatment routines. An earlier study (Walker & Hops, 1973) also documents increases in number of initiations and range of contacts, but failed to detect an increase in the mean duration of interchanges. Therefore, it seems likely that many of the effects of the present intervention will persist.

There is also rather encouraging evidence which shows maintenance and generalization effects of treatment used with out-of-control children (Walker & Buckley, 1972; Walker & Hops, 1974). In one study (Walker, Hops, & Johnson, 1975), gains made in an experimental classroom did not dissipate

when children were returned to their regular classes and followed up during the next school year. Still, persistence of effects has been the exception rather than the rule (Kazdin & Bootzin, 1972; O'Leary & Drabman, 1971). The generalization issue is one which must be more carefully considered during the decade ahead. Regrettably, the abbreviated follow-up reported here provides no further insight into the long-term impact of behavior modification procedures.

A second problem concerns the process of selecting target children. As mentioned earlier, the contract between cooperating school commissions and the experimenter precluded direct testing or intervention of any sort, and also prohibited access to a student's personal file. This arrangement enabled the research to commence several months earlier than it would have otherwise. Unfortunately, the absence of psychological testing or records of academic performance may have permitted retarded, brain-damaged, or learning disabled children to be admitted to the sample. Despite this possibility, there are three factors which would indicate that such cases were not designated as target children: (a) the sample was drawn exclusively from regular classes; (b) the study began four months after the onset of the school year, a reasonable period of time for extremely handicapped children to be identified and referred elsewhere; (c) teachers submitted global ratings of target children's academic performance; in no case was a selected pupil failing, although they tended to be below average.

The present study is hampered by another deficiency. There is reason to believe that the sample of teachers was not representative. Participants in experimental groups displayed an unusual degree of motivation by identifying and referring problem children, agreeing to classroom observations, consistently

attending training sessions (albeit for credit), supplying data, and carrying out treatment programs. Their willingness to work was, in part, controlled for by selecting control subjects who were also highly motivated, as evidenced by the fact that they were currently receiving or had recently completed some other form of supplementary instruction. Still, the results reported herein may not be generalizable to all, or even most, elementary school teachers.

There is, however, good reason to believe that many teachers would avail themselves of a behavior modification practicum. All 40 members of groups E_1-E_4 completed an evaluation of the experience. When asked to compare the present training with other post-graduate education courses, all but one rated this program "above average" and about two-thirds of these considered it "much better." When asked whether they would recommend such a program to other teachers, all but one said they would recommend it "highly." Of the positive respondents, 40% described it as indispensible. Thirteen teachers formed a committee to promote the introduction of a similar training program within their various school commissions. They were subsequently joined by a number of principals who had, by September 1974, submitted a proposal to at least one school commission. The author also received a strong indication that behavior modification would be given greater emphasis in both degree curricula and in the continuing education program at McGill University.

In summary, six original contributions were made: (a) to the author's knowledge, a comparable design allowing for intra-group (intra-subject), inter-group, and normative comparisons had not appeared in the operant literature; (b) there had been no other study of withdrawn children which utilized a matched, untreated control group; (c) this was the first attempt to identify via naturalistic observation non-peer-oriented dependent variables (e.g.,

"looking around, "self-stimulation," "volunteering") which discriminated withdrawn children from their peers; (d) this was the first formal report where observation data were transformed into standard score composites;[25] (e) this was the first naturalistic test for a tracking effect; (f) this study marked the initial attempt to utilize a demand baseline procedure in applied operant outcome research.

---

[25] Observation data gathered on the Social Learning Project at the Oregon Research Institute have recently been reanalyzed following standard score transformation.

Bibliography

Adams, S. Interaction between individual interview therapy and treatment amenability in older youth authority wards. In: Inquiries concerning kinds of treatments for kinds of delinquents. Sacramento, California: California Board of Corrections, 1961. Pp. 27-44.

Aichorn, A. Wayward youth. New York: Viking Press, 1935.

Allen, K. E., Hart, B., Buell, J. S., Harris, F. R., & Wolf, M. M. Effects of social reinforcement on isolate behavior of a nursery school child. Child Development, 1964, 35, 511-518.

Andrews, J. K. The results of a pilot program to train teachers in the classroom application of behavior modification techniques. Journal of School Psychology, 1970, 8, 37-42.

Ayllon, T., & Azrin, N. H. The measurement and reinforcement of behavior of psychotics. Journal of Experimental Analysis of Behavior, 1965, 8, 357-383.

Baer, D. M., & Wolf, M. M. The entry into natural communities of reinforcement. In R. Ulrich, T. Stachnik, & J. Mabry (Eds.), Control of human behavior. Glenview, Illinois: Scott Foresman, 1970. Pp. 319-324.

Baker, B. L., Heifetz, L., & Pasick, R. Timberlane Project: Introduction of behavior modification for retarded and non-retarded children in a rural school district. Final Report, New England Center for Continuing Education. Durham, New Hampshire, 1973.

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

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

Barker, R. G., Wright, H. F., Barker, L. S., & Schoggen, M. Specimen records of American and English children. Lawrence, Kansas: University of Kansas Press, 1961.

Barrish, H. H., Saunders, M., & Wolf, M. M. Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. Journal of Applied Behavior Analysis, 1969, 2, 119-124.

Baumrind, D. Naturalistic assessment of parent-child interaction. Paper presented at the Conference on Research Methodology of Parent-Child Interaction, Syracuse, New York, 1967.

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

Becker, W. C. The relationship of factors in parental ratings of self and each other to the behavior of kindergarten children as rated by mothers, fathers, and teachers. Journal of Consulting Psychology, 1960, 24, 507-527.

Becker, W. C., Engelman, S., & Thomas, D. R. Teaching: A course in applied psychology. Chicago, Illinois: Science Research Associates, 1971.

Bell, R. Q., Waldrop, M. F., & Weller, G. M. A rating system for the assessment of hyperactive and withdrawn children in preschool samples. American Journal of Orthopsychiatry, 1972, 42, 23-34.

Bolstad, O. D. The relationship between teachers' assessment of students and the students' actual behavior in the classroom. Unpublished doctoral dissertation, University of Oregon, 1974.

Box, G. E., & Tiao, G. C. A change in level of non-stationary time series. Biometrika, 1965, 52, 181-192.

Boyd, R. D., & DeVault, M. V. The observation and recording of behavior. In L. McLean, R. D. Bock, M. V. DeVault, D. L. Meyer, & E. B. Page (Eds.), Review of educational research. Vol. 36. Washington, D. C.: American Educational Research Assoc., 1966. Pp. 529-551.

Buckle, D., & Lebovici, S. The child guidance centers. Geneva: World Health Organization, 1960.

Buell, J., Stoddard, P., Harris, F. R., & Baer, D. M. Collateral social development accompanying reinforcement of outdoor play in a preschool child. Journal of Applied Behavior Analysis, 1968, 1, 167-173.

Calhoun, J., & Koenig, K. P. Classroom modification of elective mutism. Behavior Therapy, 1973, 4, 700-702.

Camp, B. W., & Zimet, S. G. The relationship of teacher rating scales to behavior observations and reading achievement of first grade children. Journal of Special Education, 1974, 8, 353-359.

Clement, P. W., & Milne, D. C. Group play therapy and tangible reinforcers used to modify the behavior of eight-year-old boys. Behavior Research and Therapy, 1973, 4, 700-702.

Cobb, J. A. Survival skills and first grade academic achievement. Report No. 1, Center at Oregon for Research in the Behavioral Education of the Handicapped, University of Oregon, 1970.

Cobb, J. A. Relationship of discrete classroom behavior to fourth-grade academic achievement. Journal of Educational Psychology, 1972, 63 74-80.

Cobb, J. A., & Hops, H. Effects of academic survival skill training on low-achieving first graders. Journal of Educational Research, 1973, 67, 74-80.

Cohen, A. R. Attitude change and social influence. New York: Basic Books, 1964.

Collins, R. C. The treatment of disruptive behavior problems by employment of a partial-milieu consistency program. Unpublished doctoral dissertation, University of Oregon, 1966.

Conners, C. K. Symptom patterns in hyperkinetic, neurotic, and normal children. Child Development, 1970, 41, 667-682.

Conrad, R. D., Deck, J. L., & Williams, C. Use of stimulus fading procedures in the treatment of situation specific mutism: A case study. Behavior Therapy and Experimental Psychiatry, 1974, 5, 99-100.

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

Cross, H. J. The outcome of psychotherapy: A selected analysis of research findings. Journal of Consulting Psychology, 1964, 25, 413-418.

Crow, W. J. The effect of training upon accuracy and variability in interpersonal perception. Journal of Abnormal Social Psychology, 1957, 55, 355-359.

DeMaster, B. L. Effects of differing amounts of feedback and methods of assessment on reliability of data collected by pairs of observers. Unpublished Master's thesis, University of Wisconsin, 1971.

Douglas, J. W. B., Lawson, A., Cooper, J. E., & Cooper, E. Family interaction and the activities of young children. Journal of Child Psychology and Psychiatry, 1968, 9, 157-171.

Eron, L. D., Banta, T. J., Walder, L. O., & Laulicht, J. H. Comparison of data obtained from mothers and fathers on child-rearing practices and their relation to child aggression. Child Development, 1961, 32, 457-472.

Eyberg, S. M., & Johnson, S. M. Multiple assessment of behavior modification with families: Effects of contingency contracting and order of treated problems. Journal of Consulting and Clinical Psychology, 1974, 42, 594-606.

Eysenck, H. J. The effects of psychotherapy: An evaluation. Journal of Consulting Psychology, 1952, 16, 319-324.

Festinger, L. A theory of cognitive dissonance. Stanford, California: Stanford University Press, 1957.

Festinger, L. Behavioral support for opinion change. Public Opinion Quarterly, 1964, 28, 404-417.

Fitzsimmons, S., Cheever, J., Leonard, E., & Macunovich, D. School failures: Now and tomorrow. Developmental Psychology, 1969, 1, 134-146.

Fixsen, D. L., Phillips, E. L., & Wolf, M. M. Achievement Place: The reliability of self-reporting and peer-reporting and their effects on behavior. Journal of Applied Behavior Analysis, 1972, 5, 19-30.

Glass, G. V., Willson, V. L., & Gottman, J. M. Time series analysis in the behavioral sciences. Boulder, Colorado: Laboratory of Educational Research, University of Colorado, 1973.

Goodenough, F. L. Inter-relationships in the behavior of young children. Child Development, 1930, 1, 29-47.

Guerney, B. G., Shapiro, E. B., & Stover, L. Parental perceptions of maladjusted children: Agreement between parents, and relation to mother-child interaction. Journal of Genetic Psychology, 1968, 113, 215-225.

Gussow, Z. The observer-observed relationship as information about structure in small group research. Psychiatry, 1964, 27, 230-247.

Hagen, R. L., Craighead, W. E., & Paul, G. L. Staff reactivity to evaluative behavioral observations. Behavior Therapy, 1975, 6, 201-205.

Haggard, E. A., Brekstad, A., & Skard, A. G. On the reliability of the amammestic interview. Journal of Abnormal and Social Psychology, 1960, 61, 311-318.

Hall, R. V. Training teachers in classroom use of contingency management. Educational Technology, 1971, 11, 33-38.

Halleck, S. L. Psychiatry and the dilemmas of crime: A study of causes, punishment and treatment. New York: Harper and Row, 1967.

Halverson, C. F., Jr., & Waldrop, M. F. Relations between preschool barrier behaviors and early school age measures of coping, imagination, and verbal development. Developmental Psychology, 1974, 10, 716-720.

Harmatz, M. G., Mendelsohn, R., & Glassman, M. Behavioral observation in the study of schizophrenia. Paper presented at the meeting of the American Psychological Association, Montreal, Quebec, Canada, 1973.

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

Harris, F. R., Wolf, M. M., & Baer, D. M. Effects of adult social reinforcement on child behavior. Young Children, 1964, 20, 8-17.

Hart, B. M., Reynolds, N. J., Baer, D. M., Brawley, E. R., & Harris, F. R. Effect of contingent and non-contingent social reinforcement on the cooperative play of a preschool child. Journal of Applied Behavior Analysis, 1968, 1, 73-76.

Heifitz, L. Behavior Vignettes Test: School version. Unpublished manuscript, Harvard University, 1972.

Hendriks, A. F. C. J. Reported vs. observed deviancy. Unpublished doctoral dissertation, University of Nijmegen, The Netherlands, 1972.

Herbert, E. W., & Baer, D. M. Training parents as behavior modifiers: Self-recording of contingent attention. Journal of Applied Behavior Analysis, 1972, 5, 139-149.

Heyns, R., & Lippitt, R. Systematic observational techniques. In G. Lindzey (Ed.), Handbook of social psychology. Vol. I. Reading, Massachusetts: Addison-Wesley, 1954.

Hines, P. A. How adults perceive children: The effect of behavior tracking and expected deviance on teachers' impressions of a child. Unpublished doctoral dissertation, University of Oregon, 1974.

Hjelle, L. A. Accuracy of personality and social judgments as functions of familiarity. Psychological Reports, 1968, 22, 311-319.

Honig, A. S., Tannenbaum, J., & Caldwell, B. Maternal behavior in verbal report and in laboratory observation. Paper presented at the meeting of the American Psychological Association, San Francisco, 1968.

Hops, H., & Cobb, J. A. Survival skills in the educational setting: Their implications for research and intervention. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology concepts and practice. Champaign, Illinois: Research Press, 1973. Pp. 193-209.

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

Johnson, S. M., Christianson, A., & Bellamy, G. T. Evaluation of family intervention through unobtrusive audio recordings: Experiences in bugging children. Unpublished manuscript, University of Oregon, 1974.

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

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, 1972.

Jones, R. R. Observation by telephone: An economical behavior sampling technique. Oregon Research Institute Technical Report, 1974, 14, No. 1.

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

Jones, R. R., Reid, J. B., & Patterson, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment. Vol. 3. San Francisco: Jossey-Bass, 1975. Pp. 42-95.

Jones, R. R., Vaught, R. S., & Reid, J. B. Time series analysis as a substitute for single subject analysis of variance designs. Paper presented at the American Psychological Association Convention, Montreal, Quebec, Canada, 1973.

Kale, F. D., & Toler, H. C., Jr. Modification of preschool isolate behavior: A case study. Journal of Applied Behavior Analysis, 1970, 3, 309-314.

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

Kaufman, K. F., & O'Leary, K. D. Reward, cost, and self-evaluation procedures for disruptive adolescents in a psychiatric hospital school. Journal of Applied Behavior Analysis, 1972, 5, 293-309.

Kazdin, A., & Bootzin, R. The token economy: An evaluative review. Journal of Applied Behavior Analysis, 1972, 5, 343-373.

Kent, R. N. Observer presence as an influence on teacher and child behavior in a classroom setting: A replication. State University of New York at Stony Brook, in preparation.

Kent, R. N., Fisher, J. B., & O'Leary, K. D. Observer presence as an influence on teacher and child behavior in a classroom setting. Unpublished manuscript, State University of New York at Stony Brook, 1974.

Kerlinger, F. N. Foundations of behavioral research: Educational and psychological inquiry. New York: Holt, Rinehart & Winston, 1964.

Kohn, M. The child as a determinant of his peers' approach to him. Journal of Genetic Psychology, 1966, 109, 91-100.

Krumboltz, J. D., & Goodwin, D. L. Increasing task-oriented behavior: An experimental evaluation of training teachers in reinforcement techniques. Final report. U.S. Office of Education, 1966.

LaPiere, R. T. Attitudes vs. actions. Social Forces, 1934, 13, 230-237.

Lapouse, R., & Monk, M. M. An epidemiological study of behavior characteristics in children. American Journal of Public Health, 1958, 48, 1134-1144.

Lazarus, A. A., & Davison, G. C. Clinical innovation in research and practice. In A. E. Bergin & S. L. Garfield (Eds.), Handbook of psychotherapy and behavior change. New York: Wiley, 1971. Pp. 196-213.

Leslie, D. G. The effects of systematic observation on adults' perception of behavior change in children. Unpublished doctoral dissertation, University of Oregon, 1975.

Lewinsohn, P. M. Clinical and theoretical aspects of depression. Paper presented at the Georgia Symposium in Experimental Clinical Psychology, 1972.

Lobitz, G., & Johnson, S. M. Normal versus deviant children: A multimethod comparison. Paper presented at the meeting of the Western Psychological Association, San Francisco, 1974.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Maccoby, E., & Masters, J. C. Attachment and dependency. In P. Mussen (Ed.), Carmichael's manual of child psychology. Vol. 2. New York: Wiley, 1970.

Mash, E. J., & Hedley, J. Observer effect as a function of prior history of social interaction. Unpublished manuscript, University of Calgary, 1974.

Masling, J., & Stern, G. Effect of the observer in the classroom. Journal of Educational Psychology, 1969, 60, 351-354.

McCord, J., & McCord, W. Cultural stereotypes and the validity of interviews for research in child development. Child Development, 1961, 32, 171-185.

McKinney, J. D., Mason, J., Perkerson, K., & Clifford, M. Relationship between classroom behavior and academic achievement. Journal of Educational Psychology, 1975, 67, 198-203.

Medland, M. B., & Stachnik, J. G. Good behavior game: A replication and systematic analysis. Journal of Applied Behavior Analysis, 1972, 5, 45-51.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand-McNally, 1963. Pp. 247-328.

Mercatoris, M., & Craighead, W. E. The effects of non-participant observation on teacher and pupil classroom behavior. Journal of Educational Psychology, 1974, 66, 512-519.

Meyer, H. J., Borgatta, E. F., & Jones, W. C. Girls at Vocational High: An experiment in social work intervention. New York: Russell Sage Foundation, 1965.

Michael, C. M., Morris, D. P., & Soroker, M. A. Follow-up studies of shy, withdrawn children. II: Relative incidence of schizophrenia. American Journal of Orthopsychiatry, 1957, 27, 331-337.

Milby, J. B., Jr. Modification of extreme social isolation by contingent social reinforcement. Journal of Applied Behavior Analysis, 1970, 3, 149-152.

Miller, L. C., Hampe, E., Barrett, C. L., & Nobel, H. Children's deviant behavior within the general population. Journal of Consulting and Clinical Psychology, 1971, 37, 16-22.

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

Morris, D. P., Soroker, E., & Burruss, G. Follow-up studies of shy, withdrawn children. American Journal of Orthopsychiatry, 1954, 4, 743-754.

Murphy, L. B. Social behavior and child personality. New York: Columbia University Press, 1937.

Novick, J. Rosenfeld, E., Block, A. A., & Davidson, D. Ascertaining deviant behavior in children. Journal of Consulting Psychology, 1966, 30, 230-238.

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

O'Connor, R. D. Relative efficacy of modeling, shaping and the combined procedures for modification of social withdrawal. Journal of Abnormal Psychology, 1972, 79, 327-334.

O'Leary, K. D., & Drabman, R. Token reinforcement programs in the classroom. Psychological Bulletin, 1971, 75, 379-398.

O'Leary, K. D., & Kent, R. Behavior modification for social action: Research tactics and problems. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology concepts and practice. Champaign, Illinois: Research Press, 1973.

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

O'Leary, K. D., & O'Leary, S. Classroom management: The successful use of behavior modification. New York: Pergamon, 1972.

Orne, M. T. On the social psychology of the psychological experiment. American Psychologist, 1962, 17, 776-783.

Patterson, G. R. The aggressive child: Victim and architect of a coercive system. In L. A. Hamerlynck, E. J. Mash, & L. C. Handy (Eds.), Behavior modification and families. I. Theory and research. II. Applications and developments. New York: Brunner and Mazell, 1975.

Patterson, G. R., & Cobb, J. A. Stimulus control for classes of noxious behavior. In J. F. Knutson (Ed.), The control of aggression: Implications from basic research. Chicago: Aldine, 1973. Pp. 144-199.

Patterson, G. R., Cobb, J. A., & Ray, R. S. Direct intervention in the classroom: A set of procedures for the aggressive child. In F. W. Clark, D. R. Evans, & L. A. Hamerlynck (Eds.), Implementing behavioral programs for schools and clinics. Champaign, Illinois: Research Press, 1972. Pp. 151-201.

Patterson, G. R., Cobb, J. A., & Ray, R. S. A social engineering technology for retraining the families of aggressive boys. In H. Adams & I. P. Unikel (Eds.), Issues and trends in behavior therapy. Springfield, Illinois: Chas. C. Thomas, 1973. Pp. 139-224.

Patterson, G. R., Shaw, D. A., & Ebner, M. J. Teachers, peers, and parents as agents of change in the classroom. In F. A. M. Benson (Ed.), Modifying deviant social behaviors in various classroom settings. Eugene, Oregon: University of Oregon, 1969. No. 1. Pp. 13-47.

Paul, G. L. Insight versus desensitization in psychotherapy. Stanford, California: Stanford University Press, 1966.

Paul, G. L. Extraversion, emotionality and physiological response to relaxation training and hypnotic suggestion. International Journal of Clinical and Experimental Hypnosis, 1969, 17, 89-98. (a)

Paul, G. L. Outcome of systematic desensitization. II. Controlled investigations of individual treatment, technique variations and current status. In C. M. Franks (Ed.), Behavior therapy: Appraisal and status. New York: McGraw-Hill, 1969. Pp. 105-159. (b)

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

Phillips, E. L. Achievement Place: Token reinforcement procedures in a home-style rehabilitation setting for pre-delinquent boys. Journal of Applied Behavior Analysis, 1968, 1, 213-223.

Powers, E., & Witmer, H. An experiment in the prevention of delinquency: The Cambridge-Sommerville youth study. New York: Columbia University Press, 1951.

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

Rapp, D. W. Detection of observer bias in the written record. Unpublished manuscript, University of Georgia, 1965.

Redl, F., & Wineman, D. Children who hate. Glencoe, Illinois: The Free Press, 1951.

Redl, F., & Wineman, D. Controls from within: Techniques for the treatment of the aggressive child. Glencoe, Illinois: The Free Press, 1952.

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

Reid, J. B. The relationship between complexity of observer protocols and observer agreement. Paper presented at the meeting of the American Psychological Association, Montreal, Quebec, Canada, 1973.

Reid, J. B., & DeMaster, B. The efficacy of the spot-check procedure in maintaining the reliability of data collected by observers in quasi-natural settings: Two pilot studies. Oregon Research Institute Research Bulletin, 1972, 12, No. 8.

Reid, J. B., Hawkins, N., Keutzer, C., McNeal, S. A., Phelps, R. E., Reid, K. M., & Mees, H. L. A marathon behaviour modification of a selectively mute child. Journal of Child Psychology and Psychiatry, 1967, 8, 27-30.

Reid, J. B., & Patterson, G. R. Follow-up analyses of a behavioral treatment program for boys with conduct problems: A reply to Kent. Journal of Consulting and Clinical Psychology, 1975, in press.

Robbins, L. C. The accuracy of parental recall of aspects of child development and of child rearing practices. Journal of Abnormal and Social Psychology, 1963, 66, 261-270.

Robbins, L. C. Deviant children grown up: A sociological and psychiatric study of sociopathic personality. Baltimore, Maryland: Williams & Wilkins, 1966.

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

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

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal and R. L. Rosnow (Eds.), Artifact in behavioral research. New York: Academic Press, 1969. Pp. 181-277.

Ross, D. M., Ross, S. A., & Evans, P. A. The modification of extreme social withdrawal by modeling with guided participation. Journal of Behavior Therapy and Experimental Psychiatry, 1971, 2, 273-279.

Rutter, M., & Graham, P. Psychiatric disorders in 10 and 11 year old children. Proceedings of the Royal Society of Medicine, 1965, 59, 382-387.

Schectman, A. Psychiatric symptoms observed in normal and disturbed children. Journal of Clinical Psychology, 1970, 26, 38-41.

Schnelle, J. F. A brief report on invalidity of parent evaluations of behavior change. Journal of Applied Behavior Analysis, 1974, 7, 341-343.

Schoggen, P. A study in psychological ecology: Structural properties of children's behavior based on sixteen day-long specimen records. Unpublished doctoral dissertation, University of Kansas, 1954.

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

Sears, R. R. Comparison of interviews with questionnaires for measuring mothers' attitudes toward sex and aggression. Journal of Personality and Social Psychology, 1965, 2, 37-44.

Selltiz, C., Jahoda, M., Deutsch, M., & Cook, S. W. Research methods in social relations. New York: Holt, Rinehart & Winston, 1959.

Shaw, D. A. Family maintenance schedules for deviant behaviors. Unpublished doctoral dissertation, University of Oregon, 1971.

Shepherd, M., Oppenheim, A. N., & Mitchell, S. Childhood behavior disorders and the child guidance clinic: An epidemiological study. Journal of Child Psychology and Psychiatry, 1966, 7, 39-52.

Sines, J. O., Paulker, J. D., Sines, L. K., & Owen, D. R. Identification of clinically relevant dimensions of children's behavior. Journal of Consulting and Clinical Psychology, 1969, 33, 728-734.

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

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

Sobell, L. C., Sobell, M. B., & Christelman. The myth of "one drink." Behavior Research and Therapy, 1972, 10, 119-123.

Speer, D. C. Behavior problem checklist (Peterson-Quay): Baseline data from parents of child guidance and non-clinic children. Journal of Consulting and Clinical Psychology, 1971, 36, 221-228.

Spivack, G., & Swift, M. Classroom behavior of children: A critical review of teacher administered behavior rating scales. Journal of Special Education, 1973, 7, 55-89.

Sprague, R. L., Christiansen, D. E., & Werry, J. S. Experimental psychology and stimulant drugs. Paper presented at the Symposium, "The Clinical Use of Stimulant Drugs in Children." Key Biscayne, Florida, 1972.

Stuart, R. B., & Davis, B. Slim chance in a fat world: Behavioral control of obesity. Champaign, Illinois: Research Press, 1972.

Surratt, P. R., Ulrich, R. E., & Hawkins, R. P. An elementary student as a behavioral engineer. Journal of Applied Behavior Analysis, 1969, 2, 65-72.

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

Teuber, H. L., & Powers, E. Evaluating therapy in a delinquency prevention program. Psychiatric Treatment, 1953, 21, 138-147.

Tharp, R. G., & Wetzel, R. J. Behavior modification in the natural environment. New York: Academic Press, 1969.

Tharp, R. G., Wetzel, R. J., & Thorne, G. Behavioral research report: Final report, Office of Juvenile Delinquency and Youth Development, Health, Education, and Welfare Grants #65023 and #66020, 1968.

Vasey, W. Implications for social work education. In G. E. Brown (Ed.), The multi-problem dilemma. Metuchen, New Jersey: Scarecrow Press, 1968, Pp. 32-46.

Vernon, P. E. Personality assessment: A critical survey. New York: Wiley, 1964.

Vroom, V. H. Work and motivation. New York: Wiley, 1964.

Wahler, R. G. Setting generality: Some specific and general effects of child behavior therapy. Journal of Applied Behavior Analysis, 1969, 2, 239-246.

Wahler, R. G., & Leske, G. Accurate and inaccurate observer summary reports. Journal of Nervous and Mental Disease, 1973, 156, 386-394.

Waldrop, M. F., & Halverson, C. F., Jr. Intensive and extensive peer behavior: Longitudinal and cross-sectional analysis. Child Development, 1975, 46, 19-26.

Walker, H. M. Walker Problem Behavior Identification Checklist. Los Angeles: Western Psychological Services, 1970.

Walker, H. M. Early identification and assessment of behaviorally handicapped children in the primary grades. Report No. 2, Center at Oregon for Research in the Behavioral Education of the Handicapped, University of Oregon, 1971.

Walker, H. M., & Buckley, N. K. Programming generalization and maintenance of treatment effects across time and across settings. Journal of Applied Behavior Analysis, 1972, 3, 209-224.

Walker, H. M., & Hops, H. The use of group and individual reinforcement contingencies in the modification of social withdrawal. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology concepts and practice. Champaign, Illinois: Research Press, 1973.

Walker, H. M., & Hops, H. A normative model for evaluating generalization and maintenance of treatment effects. Report No. 12, Center at Oregon for Research in the Behavioral Education of the Handicapped, University of Oregon, 1974.

Walker, H. M., & Hops, H. Use of normative peer data as a standard for evaluating classroom treatment effects. Unpublished manuscript, Center at Oregon for Research in the Behavioral Education of the Handicapped, University of Oregon, 1975.

Walker, H. M., Hops, H., Greenwood, C. R., & Todd, N. Analysis and modification of social withdrawal within an experimental class setting. Unpublished manuscript, Center at Oregon for Research in the Behavioral Education of the Handicapped, University of Oregon, 1975.

Walker, H. M., Hops, H., & Johnson, S. M. Generalization and maintenance of classroom treatment effects. Behavior Therapy, 1975, 6, 188-200.

Walker, H. I., & Gilmore, S. K. Placebo versus social learning effects in parent training procedures designed to alter the behaviors of aggressive boys. Behavior Therapy, 1973, 4, 361-377.

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

Weeks, H. A. Youthful offenders at Highfields. Ann Arbor, Michigan: University of Michigan Press, 1958.

Weinrott, M. R., Walker, H. M., & Hops, H. The influence of observer presence on classroom behavior. Oregon Research Institute, in preparation.

Werry, J. S., & Quay, H. C. A method of observing classroom behavior of emotionally disturbed children. Exceptional Children, 1968, 34, 389.

Werry, J. S., & Quay, H. C. Observing the classroom behavior of elementary classroom children. Exceptional Children, 1969, 35, 461-470.

White, G. D. The effects of observer presence on mother and child behavior. Unpublished doctoral dissertation, University of Oregon, 1973.

Whitman, T. L., Mecurio, J. R., & Caponigri, V. Development of social responses in two severely retarded children. Journal of Applied Behavior Analysis, 1970, 3, 133-138.

Wickman, E. K. Children's behavior and teachers' attitudes. New York: Oxford University Press, 1928.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Massachusetts: Addison-Wesley, 1973.

Wilchesky, M. The effects of over-training on observer reliability: A pilot study. Unpublished Honour's Thesis, McGill University, 1974.

Willems, E. P., & Raush, H. L. Naturalistic viewpoints in psychological research. New York: Holt, Rinehart & Winston, 1969.

Williams, J. G., Barlow, D. H., & Agras, W. S. Behavioral measurement of severe depression. Archives of General Psychiatry, 1972, 27, 330-333.

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

Wolff, S. Behavioral characteristics of primary school children referred to a psychiatric department. British Journal of Psychiatry, 1967, 113, 885-893.

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

Wright, J. C. G. The relative efficacy of systematic desensitization and behavioral training in the modification of university quiz section participation difficulties. Unpublished doctoral dissertation, University of Wisconsin, 1972.

Yarrow, M. R., Campbell, J. D., & Burton, R. V. Reliability of maternal retrospection: A preliminary report. Family Process, 1964, 3, 207-218.

Zax, M., Cowen, E. L., Rappaport, J., Beach, D., & Laird, J. Follow-up study of children identified as emotionally disturbed. Journal of Consulting Psychology, 1968, 32, 369-373.

APPENDIX A

Manual for Coding Interactions in the Classroom Setting[1]

In order to achieve meaningful and reliable results when recording behavior, many conditions must be fulfilled. Foremost among these is the acquisition of a skill which requires a high degree of motivation and alertness in addition to an ability to make on-the-spot discriminations. It is also generally recognized that reliable ratings of behavior can be obtained only to the extent that the code categories are clearly defined, behaviorally anchored, and involve a minimum of inference on the part of observers.

This manual is a guide to be used in preparing to observe behaviors occurring in the school setting. The coding system has been developed to provide a precise record of behavioral rates and interaction patterns.

The observer will look at the target subject (i.e., selected student) and each same-sex peer in alternating six-second intervals, i.e., subject, peer; subject, peer; subject, peer; etc. The observer will code the behavior displayed by placing a mark (—) in the appropriate box on the computer scanner sheet (see sample on last page). If there is a response to the behavior by another person (teacher or peer) which can be discerned by the target subject, the response is to be coded within the same six-second interval but in the row of boxes just below that used for the observed child. Each double row of boxes represents six seconds, the first row is reserved for the behavior and the second for the response. If there is no response, and often this will be the

---

[1]Adapted from Patterson, G. R., Cobb, J. A., & Ray, R. S. Manual for coding discrete behaviors in the school setting. Oregon Research Institute, 1971.

case, leave the second row of the interval blank. If the response is displayed by the teacher, then an additional mark should be placed in the box labeled "T" in the first column. If the response is displayed by a peer or peers, the box labeled "T" should be left blank and only the behavior category should be coded.

Observers are cautioned against making stray marks on their coding sheets as these may be registered as valid entries by the computer. Also, the scanner sheets must not be rolled, folded, exposed to excessive moisture or otherwise altered.

An auditory pacer with earplug is provided to produce a signal every six seconds so the observer will know when to code a child's behavior. An efficient procedure for coding is to observe the child for a few seconds after the auditory signal (tone) occurs and check to see if there is a response from the environment; then code the behavior observed as well as the response; if there is no immediate response, but a response occurs before the end of the six-second interval, code that response, wait for the next auditory signal and repeat the procedure for the next person. Once all same-sex peers have been coded in the classroom, the observer will begin coding in the same order of peers on the same coding sheet as in the original sequence. Sometimes the original order will be difficult to maintain due to movement in the classroom; in these cases the observer should attempt to sample all same-sex peers, regardless of order, before returning to coding the same peer twice. If a peer leaves the room or is unobservable for other reasons, do not leave the space blank; just continue and code the next peer.

Space is provided at the top and in the left margin for entering specific information about the observation session. Group and subject numbers will be obtained from a list provided by the experimenter and will be filled in at the

top of each sheet. The experimenter will enter both the phase number and observation number at a later time. The observer should note the date and time of observation in the left margin. Each observer will have her own identification number which she will record in line one. She will also note when the observation is to serve as a reliability check by marking the box "R." Space in line one is also provided for the structure of the ongoing activity and the kind of work (group, individual or transitional) that is occurring at the time of coding. If the task is individual, then the box "I" should be marked. If the task is a group project, then the box "I" should be left blank. If the activity is transitional (between tasks), then only the "TR" box should be marked. If the lesson is structured, then the "ST" box should be marked and if it is unstructured, then no mark is needed in the "ST" box. The observer is to fill in the academic subject (e.g., reading, arithmetic, social studies) in the left margin. When changes occur in the structure or in the kind of work (group, individual, or transitional), then the coding should stop on that particular sheet, the change should be noted on the next sheet and coding should continue at the top of the new page.

If individual mark "I."
If group leave "I" blank.
If transitional mark "TR"; leave "I" blank.

If structured mark "ST."
If unstructured leave "ST" blank.

The definitions for the five categories are as follows:

Structured. The teacher has provided clear guidelines for the children to follow in carrying out tasks.

Unstructured. The guidelines for the child's behavior are vague or unclear to the observer, i.e., the students can determine what they want to do in terms of academic activity.

Group. The class is involved as one unit in academic activity, e.g., the teacher lecturing, student reciting while others listen. Also, "group" applies to activities where the class is divided into several small units such as in reading or special projects.

Individual. The majority of the students are doing work by themselves at desks, e.g., art projects are being done by each child. "Individual" applies even though the student asks for and receives help from other peers and/or teachers.

Transitional. This category should be checked when the class is between activities, e.g., waiting for recess, lining up for lunch, returning from recess, teacher has indicated reading period but has provided no directions for the next activity. As soon as the teacher gives instructions for the next activity, the "TR" category is to be omitted and either "Individual" or "Group" applies.

It is essential that only one behavior be coded for each subject. Although there will be instances where more than one behavior code is applicable, the observer should code only one. To facilitate a consistent choice of categories among observers, the codes are ordered in the manual as well as on the scanner sheets in a hierarchical fashion for appropriate and inappropriate behaviors. The observer is to go from left to right until the first applicable code category is reached; that category is to be marked and no other.

The same procedure is to be followed for picking a peer or teacher response. The rule to keep uppermost in mind regarding the choice of response is that the response is specifically directed at the subject. For example, if the student is attending to his work and a peer drops a book with a loud noice, the student's behavior is coded but not the peer's behavior as the behavior was not directed at the subject. However, if the peer dropped the book on the student's desk,

then that response would be coded.

In the following list the code defintions are applicable to both behavior of the subject and to responses from teacher and peers unless noted otherwise.

AP Approval: Used whenever a person gives a clear verbal, gestural, or physical sign of approval to another individual. "Approval" is more than attention, in that it must include some clear indication of positive interest or involvement. Examples of AP are smiles, head nods, hugs, pats on the back, awarding stars or points, repeating a correctly given andswer, and phrases such as, "That's a good boy," "Thank you," "That's right," and "That's a good job."

CO Complies: This category can be checked each time the person does what another person has requested, e.g., the teachers asks class to take out notebooks and pupil does; she asks for papers to be turned in and pupil obeys; pupil asks for pencil and teacher or peer gives him one; teacher tells class to be quiet and pupil is quiet. CO to be coded only during interval in which command or request occurred or within 12 seconds following. Not to be coded more than once per command or request for target child or peer group selection.

T+ Appropriate interaction with teacher: This category can be checked when the pupil talks or interacts with the teacher, whether in private as in independent work situations or answers questions in other situations. If the teacher is interacting with the child when the child is behaving appropriately, the response is coded T+. The reason for coding the subject's behavior and the response in the same category is the difficulty of differentiating other responses in quick or hard-to-hear verbal exchanges; of course, if other responses are appropriate, especially AP or DI, and can be clearly differentiated, they preclude the coding of T+ as a response.

P+ Appropriate interaction with peer: Coded when the pupil is interacting

with peer and is not violating classroom rules. Interaction includes verbal and non-verbal communication, e.g., talking, handing materials, working on a project with peer. The response for the peer is P+ if the peer is interacting with the subject. The main element to remember in applying this code is that an interaction is occurring or one of the persons is attempting to interact. If two students are working on a social studies project, the code is P+; if they are talking to each other or organizing a project together, the code is P+; but if the subject is simply writing and the peer is writing, then the appropriate code is AT.

VO Volunteers: Coded when a person indicates that he wants to make an academic contribution, e.g., teacher asks a question and he raises his hand. Also coded when pupil gives appropriate verbal response when teacher asks question to the class as a whole without requiring that students raise hands to be recognized.

IT Initiation to or by teacher: Pupil or teacher initiates or attempts to initiate interaction with each other, but not in conjunction with volunteering. Pupil may go to teacher's desk during independent study or raise his hand and seek assistance in solving an arithmetic problem; as a response, teacher may initiate interaction with pupil either by approaching pupil's desk, calling on pupil, etc. The important aspect to consider is that the observer does not know the content of the upcoming interaction.

LA Laugh: Used whenever a person (student or teacher) laughs in a non-humiliating way while attending to task. For example, a person makes a funny remark and other people laugh at it. However, if one of the people who heard the remark laughed in a derogatory manner at the person then that would be coded as DI. It is important to remember that smiling is not sufficient for the code

LA to be used.

AT Attend: This category is used whenever a person indicates by his behavior that he is doing what is appropriate in a school situation, e.g., he is looking at the teacher when she is presenting material to the class; he is looking at visual aids as the teacher tells about them; he has his eyes focused on his book as he does the reading assignment; he writes answers to arithmetic problems; the teacher or peer looks at the child reciting; the student is reading orally; the child is watching others during a break between tasks or periods. AT is to be coded as a response when there is a clear indication that the subject is aware that a teacher or peer is attending to him. Thus, when a child is working, and the teacher looks at him, the child must make some recognition of the attending on the teacher's part, e.g., he looks at the teacher. AT should be coded as a teacher response when the child is reading orally even if the child does not look up from his book.

PA Physical aggression: Used whenever an individually physically assaults or restrains another. Child makes a forceful movement directed at another either directly or by utilizing a material object as an extension of the hand, e.g., blocking others with arms or body, tripping, kicking, pinching, hitting, or throwing objects at another person. PA also includes destruction of other's materials or possessions even if the owner is in another area of the room, e.g., tearing or crumpling others' work, breaking crayons, misusing others' books (ripping out pages, writing in them, etc.), writing on another child or on another child's work. PA also coded when an individual grabs someone else's material in an intense, severe manner, e.g., pupil grabs book out of hands of another child or pupil grabs his own material from another child.

DI Disapproval: Used whenever the person gives clear verbal or gestural

disapproval of another person's behavior or characteristics. Shaking the head or finger or placing the finger over the lips to indicate "quiet" are examples of gestural disapproval. "I do not like that tone of voice," "You didn't finish your work on time;" "Your paper is sloppy," "You're quiet now, but why can't you be that way all the time," "I don't know what's gotten into you today," "I don't like you," are examples of verbal disapproval. In verbal statements it is essential that the statement explicitly states disapproval of the subject's attributes or displeasure at his behavior. DI should not be coded when a pupil gives an incorrect answer and the teacher gives a mild "No" and moves on to another person or question. If the teacher expresses greater displeasure by demanding that the child repeat the correct answer several times, by shaking her head, by sighing, by telling the child to pay attention, or by saying "No" in a loud or degrading manner, then DI should be coded. DI should not be used when the teacher is disapproving of the class's behavior in general, e.g., "You are all getting too loud," "You people did very poorly this afternoon."

NC Noncompliance: To be coded whenever the person does not do what is requested. This includes teacher giving instructions to entire class and the subject does not comply. The child gives a negative response or fails to respond to a command or request. Examples: teacher asks child to respond and child remains quiet; child answers back when a reply is either not acceptable or requested; teacher asks child to stop doing something and child continues; peer asks or commands subject to act and subject refuses. NC is to be coded only once per command or request during the interval when the command or request occurred or within 12 seconds following. If command or request is repeated, then NC may be coded again.

T- Inappropriate interaction with teacher: Used whenever content of

conversation is negative toward teacher by pupil or when classroom rules do not allow interaction with teacher. Examples are: "I'm tired of doing this lesson," "I won't start until you help me," "I can't do it" (before child even tries); groaning when raising hand to volunteer; calling out answer when it isn't one's turn. This category should not be used if DI or NC is appropriate.

P- Inappropriate interaction with peer: Coded whenever peer or pupil interacts with or attempts to interact with each other and classroom rules are being violated. Includes playing during a work period; touching a peer to get his attention; calling a peer from across the room; talking to peer during independent work; smiling-giggling between peers when they should be working.

HR High rate: Used to describe gross motor activity or verbal activity which is not directed toward another but which interveres with work and disrupts others. Examples are excessive fidgeting, dropping books loudly on floor, scraping chair, drumming with pencil, etc.

IL Inappropriate locale: This category is not to be used when rules allow for pupils to leave seats or places without permission and what the pupil is doing is not an infraction of other rules, e.g., a pupil goes to sharpen pencil would not be classified IL, unless he stopped and looked at others or at objects for a prolonged period of time. If IL has been coded and pupil begins to interact with others, then P- should be coded for the duration of the interaction.

SS Self-stimulation: A narrow class of events in which the person attempts to stimulate himself repetitively, i.e., the physical actions of the subject are not directed toward any apparent environmental stimulus. Differs from HR in that SS must be both repetitive and non-interfering with others. Examples of SS are: rubbing or poking oneself with either another part of the body or by using a material object; repetitive head movements; repetitive finger or hand

flapping; pulling or twisting hair; facial grimaces or twitches; scratching; swinging feet; scratching a pencil back and forth across the desk. SS should be coded only when attention to relevant activities is precluded. Note: SS also includes talking to oneself in a repetitive, humming manner during which the child is off-task.

LO Looking around: Coded when a person is looking around the room, daydreaming, looking out the window, or staring into space when an academic activity is occurring. To be coded LO a person must be looking around for the entire interval except for a momentary glance at his work. LO should also be coded when a pupil is singing or reciting as requested but is not attending to the presenting stimulis and is, instead, daydreaming or looking at something unrelated to the activity.

NA Not attend: This category is used when a person is not attending to work during individual work situations or not attending to discussion when the teacher or another student is presenting material. This category is applicable to those situations in which the subject is working but he is working on the wrong assignment. NA used when child is engaged in activities such as reading comic book, playing with hockey cards, etc., during lesson.

Following is a hypothetical situation in a school setting. The coding of each sequence is on accompanying coding sheets.

The observer has entered the classroom and will be coding the first sheet of the observation. The teacher is presenting a lesson in arithmetic to the while class:

1. The subject is looking out the window and the teacher says, "Jimmy, don't you ever pay attention to what's going on?"

2. The first male peer is looking at the teacher.

3. The subject looks at the teacher.

4. The second male peer is scratching and looking at his arm.

(1) 5. The subject talks to a peer while the teacher is still presenting the lesson. The peer talks with the subject.

(2) 6. The third male peer answers a question from the teacher. The teacher smiles and says, "Fine."

(3) 7. The subject pushes a book off his desk onto the floor. Several peers giggle.

(4) 8. The fourth male peer is rolling a ball down the aisle to his friend. The friend rolls it back.

(1) 9. The subject raises his hand in response to a question asked of the class by the teacher.

(2) 10. The fifth male peer picks up a piece of paper at the teacher's request. The teacher says, "Thank you."

(3) 11. The subject rummages through his desk while the teacher is presenting the lesson.

(4) 12. The sixth male peer is walking around the room. Several of his classmates look at him.

(1) 13. The subject looks at the teacher.

(2) 14. The seventh male peer hits the child next to him. The child hits him back.

(3) 15. The subject raises his hand as the teacher is talking. She does not look at him.

(4) 16. The eighth male peer looks at the teacher.

(1) 17. The subject still has his hand raised. The teacher asks him what he wants.

(2) 18. The first male peer looks at the teacher.

(3) 19. Subject stomps his foot on the floor. Several peers look at him.

(4) 20. With the teacher's permission, the second male peer explains the lesson to a neighbor, who responds with questions.

(1) 21. Subject stares at the child sitting next to him. The child does not respond.

(2) 22. The third male peer talks to the teacher about the lesson. She answers.

(3) 23. Subject talks to child sitting next to him. The child responds.

(4) 24. The fourth male peer looks around the room.

(1) 25. The subject is reading a comic book.

(2) 26. The teacher has told the fifth male peer to sit up straight. He still slouches in chair.

(3) 27. The subject is still reading a comic book. The teacher takes it away from him.

(4) 28. The sixth male peer says to the teacher, "That's a nice dress you're wearing." The teacher looks at the child and smiles.

BEHAVIOR CODING FORM: SCHOOL

PLEASE DO NOT MARK IN THIS SPACE

GROUP — PHASE — OBSERVER

SUBJECT — OBSERVATION

DATE
21 / 1
DAY / MO

TIME
10 30
11 30

SHEET
1/18

SUBJECT'S NAME: C...

SCHOOL: R...

ACADEMIC ACTIVITY: A...

COMMENTS

1	S	1
2	P	2
3	S	3
4	P	4
1	S	1
2	P	2
3	S	3
4	P	4
1	S	1
2	P	2
3	S	3
4	P	4
1	S	1
2	P	2
3	S	3
4	P	4
1	S	1
2	P	2
3	S	3
4	P	4
1	S	1
2	P	2
3	S	3
4	P	4
1	S	1
2	P	2
3	S	3
4	P	4

Instructions to Observers

1. After your first observation of the day, phone the Allen to find out if any of the teachers or target children you will be observing later in the day have called in absent.

2. Always check in at the office upon arrival at school. Someone will show you where the correct room is.

3. You should never talk to children while in the halls of a school building unless it is to get directions to the office or a room.

4. If, for any reason, you are unable to do an observation, please call either Mark or Marc.

5. Do not talk to teachers about the training they are receiving in connection with this project.

6. Make sure that the information at the top and along the left margin of the coding forms is filled in by the end of the day. REMEMBER, STRUCTURED-UNSTRUCTURED, GROUP-INDIVIDUAL-TRANSITIONAL MUST BE FILLED IN AS YOU GO ALONG.

7. Use a blunt pencil (HB) and make marks as long as possible without going outside of the appropriate box.

8. Do not interact with children while in the classroom.

9. Remember to have the teacher complete the rules form at the beginning of the session. If she is planning several activities, she may wish to fill out an additional form or two at the beginning. You should feel free to have her fill out additional forms whenever you are unclear as to what the rules are.

10. Don't be afraid to change your vantage point if you don't have a good view of the target child.

11. During reliability checks make certain that both observers begin a new sheet at the same interval. Do not confer with each other about the code categories themselves unless very unusual circumstances arise.

12. Weekly retraining and tests on code definitions will be held on Friday afternoons (time to be arranged).

CLASSROOM RULES

Interaction

During this activity (lesson, period, etc.) the following rules apply: (check those which apply)

___ No talking with peers permitted

___ Quiet talking about work permitted

___ Quiet talking in general permitted

___ Other (specify) _____

Movement

___ No movement from work area permitted

___ Movement permitted if necessary to get materials or information related to ongoing act'vity

___ Free movement permitted

Volunteering

___ Students must raise hands to be recognized

___ Students may volunteer information without raising hands

PLEASE KEEP IN MIND THAT WE ARE AWARE THAT RULES ARE MADE TO BE BROKEN OR CHANGED QUITE OFTEN. DON'T FEEL COMPELLED TO STICK TO THESE RULES JUST BECAUSE YOU FILLED OUT THIS FORM IN A CERTAIN WAY.

IF THE ACTIVITY CHANGES WHILE THE OBSERVER IS PRESENT, SHE MAY ASK YOU TO FILL OUT ANOTHER SHEET. IF YOU SPEND MORE THAN 15 SECONDS FILLING OUT THIS FORM THEN YOU ARE TAKING TOO MUCH TIME. PLEASE EXCUSE ANY INCONVENIENCE THIS MAY CAUSE.

Observer Checklist

_____ Coding forms

_____ Rules forms

_____ Daily Schedule

_____ Pencils

_____ Beeper and ear jack

_____ Extra Battery

_____ Directions to schools

_____ Map of Montreal

_____ Directory of teachers and target children

_____ Phone numbers of Allan Memorial (842-1251 loc. 1628) and Mark Weinrott (845-6395, 392-5894)

_____ Screwdriver

Teacher's Name_____ School_____ Grade_____

The purpose of this questionnaire is to identify some of the characteristics of a child in your class who could be described as "socially withdrawn." Please select the student who most closely fits the terms "withdrawn" or "isolated" and write his/her name here (first name only): _____. Kindly complete the following chart by checking the box which shows how often this child exhibits each of the behaviors listed as compared to his classmates.

Behavior	Never	Rarely	Ocassionally	Average for class	Slightly more than average	Quite often	Nearly always
volunteers in class							
daydreams							
protests when others hurt, tease or criticize him							
isolates himself from others							
laughs							
initiates conversation with peers							

1. Would you say that this child has relatively few friends? _____

2. Is this child frequently absent from school? _____

3. Which of the following statements best describes your assessment of this child's withdrawn condition?

_____ a. a serious problem which requires professional intervention
_____ b. a minor problem which is worthy of a casual short-term treatment program
_____ c. not really a problem but a condition which interferes somewhat with performance in or enjoyment of social activities or group work.
_____ d. not a problem at all; child suffers no discomfort or loss of opportunity; no cause for concern

4. How would you rate this child's overall academic performance (circle one):

excellent good average fair poor

RETURN TO: Teacher Training, Behavior Therapy Unit, Allen Memorial Institute, 1033 Pine Ave.,W., Montreal, P.Q. H3A 1A1

Teacher's Name_____ School_____ Grade_____

The purpose of this questionnaire is to identify some of the characteristics of a child in your class who could be described as "acting out." Please select the student who most closely fits the terms "acting out" or "aggressive" and write his/her name here (first name only): _____. Kindly complete the following chart by checking the box which shows how often this child exhibits each of the behaviors listed as compared to his classmates.

Behavior	Never	Rarely	Occasionally	Average for class	Slightly more than average	Quite often	Nearly always
readily complies when asked to perform tasks							
teases other children; tattles							
temper outbursts, explosive and unpredictable mood shifts							
seeks attention of teacher							
acts "smart"							
strikes back with aggressive behavior when teased or interfered with							
distorts the truth by making statements contrary to fact							

1. Is this child frequently absent from school? _____

2. Which of the following statements best describes your assessment of this child's "acting out"?

_____ a. a serious problem which requires professional intervention
_____ b. a minor problem which is worthy of casual short-term treatment
_____ c. not really a problem, but a condition which interferes somewhat with the development of peer relationships and the normal operation of the class
_____ d. not a problem at all; child suffers no discomfort or loss of opportunity; no cause for concern

3. How would you rate this child's overall academic performance performance (circle one):

excellent good average fair poor

Teacher's Name _____ School _____ Grade _____

The purpose of this questionnaire is to identify some of the characteristics of a child in your class who could be described as "distractible." Please select the student who most closely fits the term "distractible" and write his/her name here (first name only): _____. Kindly complete the following chart by checking the box which shows how often this child exhibits the behaviors listed as compared to his classmates.

Behavior	Never	Rarely	Occasionally	Average for class	Slightly more than average	Quite often	Nearly always
finishes things he starts							
out of seat							
disturbs others: teasing, provoking fights, interrupting							
excitable or impulsive							
restless; fidgeting							
demands must be met immediately or else easily frustrated							

1. Is this child frequently absent from school? _____

2. Which of the following statements best describes your assessment of this child's distractibility?

_____ a. a serious problem which requires professional intervention
_____ b. a minor problem which is worthy of a casual short-term treatment program
_____ c. not really a problem, but a condition which interferes somewhat with performance in or enjoyment of school activities or assignments
_____ d. not a problem at all; child suffers no interference with learning; no cause for concern

3. How would you rate this child's overall academic performance (circle one):

excellent good average fair poor

Behavior Modification in the Classroom
Situations Questionnaire (Form H)

Circle the ONE answer you feel is best. Some answers will be partially correct but one answer is the best. You should read through all the answers before responding.

1. Jim deliberately talks silly very often. A lot of times the things he says make no sense at all. In order to reduce this behavior, the teacher should:

   a. consult a speech therapist
   b. ignore him at those times and talk to him when he makes sense
   c. make fun of his silly talk
   d. scold his silly talk, and tell him to talk sense
   e. try to understand what he's saying, translate it into normal talk, and encourage him to imitate it

2. Peter often has temper tantrums in class. His teacher has been keeping a record of these tantrums for two weeks, and has found that he averages 4 or 5 tantrums a day. This two week record of behavior is:

   a. a waste of time, which would be better spent in doing something about his behavior
   b. necessary information for her to have before starting an effective program for the tantrums
   c. not needed at this point, since she has not yet done anything which would change his tantrums
   d. not very useful except in the hands of a psychologist
   e. good practice in observing and recording behavior

3. When angry, Margaret whines and kicks or hits the person nearest her. The best procedure to eliminate this behavior is:

   a. isolate her for a fixed amount of time
   b. talk with her about why she feels upset
   c. scold her
   d. isolate her immediately until she calms down
   e. ignore her, so she won't be rewarded by your attention

4. Billy is constantly out of his seat in class. A new program is introduced which gives Billy a token every time he stays in his seat for 5 minutes. Which of the following would best suggest that a token is a reward for Billy?

   a. Billy can exchange his tokens for a variety of things which he enjoys
   b. Billy stays in his seat longer on each of the next three days
   c. Billy proudly shows his tokens to teachers and visitors
   d. Billy becomes very upset when another child steals his tokens
   e. Billy trades his token at the class store for candy, which he likes

Circle the ONE answer you feel is best. Some answers will be partially correct but one answer is the best. You should read through all the answers before responding.

5. You have devised a program to reduce a child's temper tantrums. After using the program for a week, you find that the tantrums are occurring more often than before. You should:

   a. maintain the new program exactly as written for another week, and change it if there is still no improvement
   b. abandon the program, since it is clearly ineffective
   c. put the program aside, and try it again in a week or two
   d. change the program somewhat, and see if it works any better in its new version
   e. continue the program, but use different punishment

6. Andy has pushed other children in the cafeteria only once a week on the average for the past couple of weeks. Until then he had been pushing someone almost every day. What should you do now?

   a. keep rewarding him for behaving well in the cafeteria
   b. decide that he has done very well and stop giving him a reward
   c. begin working on one of his other problem behaviors
   d. a. and c. above
   e. b. and c. above

7. Roberta is a seven-year-old girl who is constantly complaining of various aches and illnesses which have no medical basis. She takes up a great deal of her teacher's time with these complaints. The best way for her teacher to handle the problem is:

   a. distract her by doing something with her that she enjoys
   b. ignore her when the situation arises
   c. ignore her when the situation arises, while making a special effort to attend to her appropriate behaviors
   d. console her and comfort her, so she will gain a feeling of being loved and will not need to use illness as an attention seeking device
   e. tell her that if she's really sick, she will have to have a shot

8. Which of the following would <u>always</u> be a reward for a child?

   a. candy
   b. praise
   c. special privileges
   d. none of the above
   e. two of the above

Circle the ONE answer you feel is best. Some answers will be partially correct but one answer is best. You should read through all the answers before responding.

9. You are teaching Debbie her colors. However, she is very inattentive and is easily distracted. She leaves the table at which you are working every minute or so, prefers to explore the room, etc. In this situation you should:

   a. try to interest her by pointing to her clothing and telling her what the different colors are
   b. do something else that she'd like to do, so that at least she might learn something, even if it isn't her colors
   c. offer her a substantial reward for working on the colors
   d. reward her for sitting quietly and listening
   e. forget it for the present, but promise her a reward if she will work with you later

10. Jay is a good reader who is reluctant to participate in his reading group. He would rather read by himself and enjoys the storybooks kept on a table in the rear of the room. A good teaching procedure would be to:

   a. tell him he will not be allowed to use the storybooks unless he participates in the reading group
   b. tell him he can read by himself for 5 minutes and then must join the group
   c. tell him if he reads with the group for 20 minutes, he can use the storybooks
   d. wait until he tires of the storybooks and then encourage him to join the group
   e. tell him you will work with him individually and gradually try to work him into the group

11. Sam's kindergarten teacher is teaching him to button his coat. A good method for teaching him would include:

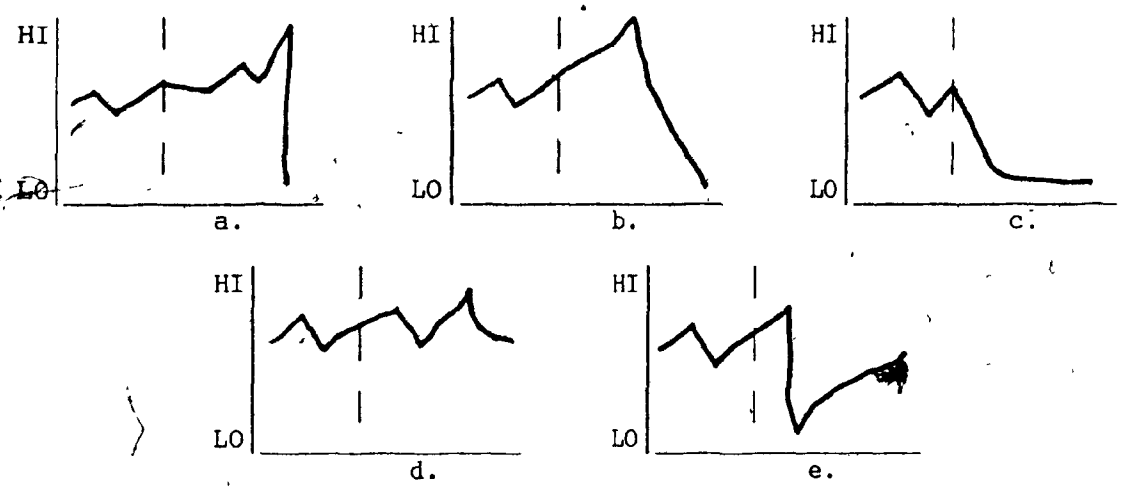
   a. starting by placing his hands on hers, while she buttons his coat
   b. using small buttons to fit his small hands
   c. teaching in reverse order the steps involved in buttoning
   d. two of the above
   e. none of the above

12. Arnie knows ten letters of the alphabet. Your toal is to teach him all of the letters. You should:

   a. start with one or two letters he doesn't know and all the ones he does know
   b. start with mainly letters he doesn't know so that he will be challenged by the difficulty of the task
   c. start with about half and half so he will succeed sometimes without being bored
   d. proceed as in c., but also reward him for successes
   e. proceed as in a., but also reward him for successes

Some answers will be partially correct, but one answer is the best. You should read through all the answers before responding.

Each of the graphs below describes a child's behavior. HI and LO can refer either to frequency of the behavior or quality of the behavior, whichever is appropriate. To the LEFT of the dashed vertical line is the behavior BEFORE a program was started to deal with it. To the RIGHT of the dashed line is the behavior AFTER the program began.



13. Which graph best describes what happens when a child begins to tire of a reward which has been used in a program to encourage some behavior of his?

_____

14. Jerry used to fool around during lessons where the material was written on the board. An eye exam revealed him to be nearsighted. Which graph best describes what happened to his inappropriate behavior after he got glasses and was placed in the front row? _____

15. If a teacher wants to discourage a certain behavior, and uses as a punishment something which is only slightly unappealing to the child, which graph best describes how that behavior changes? _____

16. Bobby has learned that you will punish her by putting her in the corner for 5 minutes if she spits. When would it be appropriate for her to have a second chance?

a. when she promises not to do it again
b. when you know she did it just to see what you would do
c. when you feel guilty about putting her in the corner
d. when you think she has forgotten what would happen if she spit
e. none of the above

Circle the ONE answer you feel is best. Some answers will be partially correct, but one answer is the best. You should read through all the answers before responding.

17. Fred is a teacher's delight. He's at or near the top of his class in all of his subjects. He is always polite and well groomed, and has never needed any discipline other than occasional scolding. Fred's all-around good behavior is most likely due to:

    a. his being the one really outstanding student who seems to appear in every class
    b. his healthy, positive psychological make-up
    c. his being an all-around good kid with a good attitude
    d. his being rewarded in some way for his behavior
    e. the fact that his parents showed a good deal of care and concern over his development

18. If you were using backward chaining to teach John Miller to print his name, what would the the FIRST letter he would print on the FIRST day of teaching?

    a. the letter J
    b. the letter R
    c. the letter N
    d. the letter M
    e. either J or M

19. Which of the following educational goals is stated in behavioral terms?

    a. to learn the multiplication tables from 1 to 9
    b. to behave appropriately during reading class
    c. to spell 15 words from a standard second grade spelling list without making more than two errors
    d. to display less hyperactive behavior
    e. b. and d. above

20. For which of the following reasons MAY a child be unable to perform a particular task:

    a. the task needs to be broken down into smaller steps
    b. the task is rewarded only every third time it is done
    c. the child has all the prerequisites for the task
    d. two of the above
    e. all of the above

Behavior Modification in the Classroom
Situations Questionnaire (Form L)

Circle the ONE answer you feel is best. Some answers will be partially correct but one answer is THE BEST. You should read through all the answers before responding.

1. Rewards are useful in teaching a child because:

   a. the child likes them
   b. the child asks for them
   c. the child exhibits certain behaviors more frequently when those behaviors are followed by the reward
   d. the child gets them for behaving well
   e. they are supposed to encourage certain behaviors which you would like to see increase

2. Barb is a new child in class. To reward her for good behavior and/or performing well in class, which of the following would be appropriate to consider:

   a. praise
   b. candy
   c. special privileges
   d. two of the above
   e. all of the above

3. Johnny has been told to stay after school because of his disruptive behavior in class. For the rest of the day, his behavior improves, and is actually quite appropriate. At the end of the day, his teacher should:

   a. allow Johnny to go home in time but discuss with him beforehand the reasons why
   b. keep him after school anyway; praise him for his good behavior
   c. let Johnny go home but send a note to his mother
   d. let Johnny go hom since he has improved
   e. let Johnny go home but substitute another punishment (for example, no recess tomorrow)

4. One of Stan's behavior problems is fighting with other children in the family. His parents have kept daily records of his fighting behavior and find that he fights less on weekends than he does during the week. This could be because:

   a. the other children have more time for him on weekends, and he doesn't have to fight to get their attention
   b. his father is home on weekends, and gives him popcorn and coke for stopping fighting
   c. on weekends the family goes on outings where there are more enjoyable things to do than fight
   d. two of the above
   e. all of the above

Circle the ONE answer you feel is best. Some answers will be partially correct, but one answer is THE BEST. You should read through all the answers before responding.

5. In language class, Penelope dawdles, babbles and giggles when asked to go to the board and identify a word. When worked with individually she can do this task well and enjoys success. What might a teacher try?

   a. have occasional individual sessions with Penelope and reward her for answering in the group sessions
   b. when she becomes inattentive, tell her to sit down and miss her turn--reward her when she does answer correctly
   c. give her as much time as she needs to answer the question, then reward her immediately
   d. prompt her until she responds correctly, and then reward her immediately
   e. do not work with her until she has been attentive for 5 minutes

6. Which of the following would <u>always</u> be a punishment for a child?

   a. isolation
   b. praise
   c. two of the above
   d. angry scolding
   e. none of the above

7. Use of an "activity reward" involves two behaviors: a behavior which the child enjoys doing; a behavior which you would like the child to perform. When such a reward is used properly:

   a. the child can engage in the behavior which he likes after he performs the behavior which you want
   b. you permit the child to engage in the behavior which he wants, but only after he promises to do the behavior which you want
   c. you permit the child to engage in the behavior he prefers, and when he tires of it, you encourage him to perform the behavior which you want
   d. the child may do the behavior he prefers for a short time, but must perform the behavior which you want before he can do any more of his preferred behavior
   e. the behavior you want and the behavior he likes are one and the same

8. Elmer is constantly getting out of his seat during his reading class. He always wanders to the same cardboard clock, which he likes to play with. Which would be the best way to decrease this inappropriate wandering:

   a. isolate him for playing with the clock
   b. minimize opportunities to leave his seat. During the period tell him if he sits down and works for 5 minutes, then he can play with the clock for 5 minutes
   c. give strong, verbal command each time he leaves his seat: "Sit down, Elmer."
   d. keep Elmer busy constantly and don't give him an opportunity to leave his seat during the entire period
   e. remove the clock

Circle the ONE answer you feel is best. Some answers will be partially correct but one answer is THE BEST. You should read through all the answers before responding.

9. Sarah has learned to match correctly the color red, when given the alternatives of red and white. You want to add another alternative. A good choice would be:

   a. orange
   b. purple
   c. brown
   d. blue
   e. cannot say

10. Ricky is a child who does not spend any time on his school work at home. His parents are encouraging him to do so, starting with 5 minutes at a time and gradually increasing it. Why would it be incorrect to reward him with candy for each small step he takes toward being able to work independently at home?

   a. because he will expect a candy reward every time he studies
   b. because it's not right to bribe him in this way
   c. because he only likes candy a little, and prefers soft drinks more
   d. because it will spoil his supper
   e. because his brother and sister learned to study without getting candy

11. To improve Jill's coordination, her kindergarten teacher has been teaching her to string beads. She has gone from being unable to string a single bead to being able to string six or seven beads in a row. In the middle of one teaching session, Jill has trouble stringing seven beads, drops them, and gets rather upset. Her teacher should:

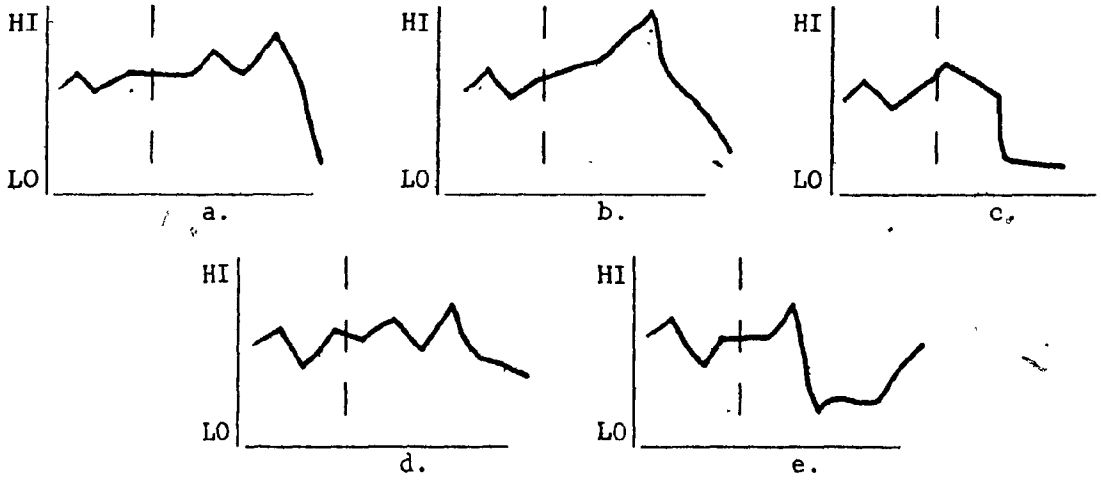
   a. end the session for the day
   b. try to soothe her and calm her, and then have her try to string the seven beads again
   c. have her string three or four beads and then end the session
   d. calm her down, reward her for doing as much as she did, and then end the session
   e. have her string three or four beads, reward her, and then end the session

12. You are using a check card in teaching a child simple counting of a few objects at a time. He has just correctly counted a small pile of blocks. In rewarding this correct response, you should:

   a. immediately have him count another small pile and thus provide him with two quick successes in a row
   b. wait a moment before giving him the check to allow him to enjoy his success
   c. offer him several checks if he can count a slightly larger pile
   d. praise him immediately and then give him a check
   e. give the check immediately and then praise him

Some answers will be partially correct, but one answer is the best. You should read through all the answers before responding.

Each of the graphs below describes a child's behavior. HI and LO can refer either to frequency of the behavior or quality of the behavior, whichever is appropriate. To the LEFT of the dashed vertical line is the behavior BEFORE a program was started to deal with it. To the RIGHT of the dashed line is the behavior AFTER the program began.



13. Which graph best describes what happens to a problem behavior (e.g., tantrums, hitting other children, playing with materials inappropriately) when an effective program is used to reduce that behavior? _____

14. Terry has trouble with multiplication. Her teacher is successfully using a program of individual work with her, very gradually increasing the difficulty of the problems which Terry can do. Which graph best describes what happens to the quality of Terry's work when her teacher too quickly increases the level of difficulty? _____

15. If a teacher wants to encourage a certain behavior, and uses as a reward something which is only slightly appealing to the child, which graph best describes how that behavior changes? _____

16. Betty had a habit of talking loudly to her neighbors during class. By rewarding her for not talking out, her teacher gradually reduced this inappropriate behavior to an acceptable level, and then discontinued the rewards. Which graph best describes what happened to Betty's talking?

_____

Circle the ONE answer you feel is best. Some answers will be partially correct but one answer is the best. You should read through all the answers before responding.

17. Chuck is a teacher's despair. He's at or near the bottom of the class in all his subjects. He is ill-mannered and poorly groomed, and has quite often needed discipline. Chuck's all-around bad behavior is most likely due to:

    a. his being the one really hopeless student who seems to appear in every class
    b. his unhealthy, negativistic psychological make-up
    c. his being an all-around bad kid with a poor attitude
    d. his being rewarded in some way for his behavior
    e. the fact that his parents showed insufficient care and concern over his development

18. Backward chaining would be an effective teaching procedure in which of the following situations:

    a. learning to identify colors
    b. learning not to answer without first raising one's hand and being recognized
    c. learning a short poem
    d. two of the above
    e. all of the above

19. Which of the following educational goals is stated in behavioral terms?

    a. to learn to do short division
    b. to behave properly during bathroom break
    c. to be less aggressive in one's behavior with one's peers
    d. to work on a task for an hour without asking more than two questions
    e. b. and c. above

20. If a child is not performing well in a particular academic task, it MAY be because:

    a. the reward for the task is obtainable elsewhere
    b. the directions are unclear
    c. the reward is contingent upon the task
    d. a. and c.
    e. a. and b.

Teacher I.D. _____

Date _____

EXPECTANCY QUESTIONNAIRE

Weather forecasting in terms of percentage (probability) is a process with which we are all familiar. For example, on Monday we may hear that there is an 80% chance of rain. On Tuesday the proability may be 20%, by Wednesday 90%. Keeping this analogy in mind, it is possible to express many aspects of our daily living in terms of probability, i.e., what are the chances of getting a parking ticket, passing an exam, etc.

In this questionnaire, we would simply like you to predict on the basis of your participation in this program what your expectations today are that the difficulties you came to work on will improve.

Circle one percentage below:

0% 10% 20% 30%	40% 50% 60% 70%	80% 90% 100%
LOW	MEDIUM	HIGH
Low probability that difficulties will improve	Medium probability that difficulties will improve	High probability that difficulties will improve

APPENDIX B

Rationale and Procedure for Standard Score Transformation of Behavioral Observation Data

Composites or clusters of responses are frequently formed from multi-category behavioral coding schemes. Composites may be constructed empirically (Patterson, 1975; Patterson & Cobb, 1973) or arbitrarily (Williams, Barlow, & Agras, 1972). In either case, one is attempting to isolate those discrete behaviors which produce a high multiple correlation with, and constitute a global construct representative of, socially relevant criterion variables.

To illustrate, a 29-category behavior coding system was found to contain 14 noxious behaviors as determined by mothers' ratings of aversiveness (Jones, Reid, & Patterson, 1975). These 14 behaviors are: command negative, cry, disapproval, dependency, destructive, high rate, humiliate, ignore, noncomply, negativism, physical negative, tease, whine, and yell. Subsequent analyses of the deviant categories yielded two distinct clusters: (a) hostility, which includes disapproval, negativism, humiliate, ignore, and whine, and (b) social aggression, which consists of physical negative and tease. The classes differ with respect to their functional control over specific responses of different family members (Patterson, 1975). In addition to "total deviant behavior," "hostility," and "social aggression," a dependent variable labeled "targeted deviant behavior" encompasses only those specific behaviors which were directly treated in a family intervention program.

A second illustration is provided by Cobb (1970), who arbitrarily designated 15 behavior categories as either appropriate or inappropriate academic survival skills. Post hoc analysis showed that a cluster of three classroom behaviors--attending, volunteering, and look around--contributed the major

portion of variance to performance in reading of first grade pupils.

Most operant investigators continue to use rather gross variables encompassing a fairly large number of categories. These include "inappropriate" (Walker & Buckley, 1972), "depressed" (Lewinsohn, 1972), "schizophrenic" (Harmatz, Mendelsohn, & Glassman, 1973), and "disruptive" (Kaufman & O'Leary, 1972). Each of these variables, however global, is derived by combining scores for a number of discrete responses. This discussion pertains to the manner in which behavior rates, frequencies, or durations for two or more categories may be synthesized into a more relevant, illustrative, and predictive construct.

Typically, behavioral observations involve time-sampling of specific responses and recording of events within prescribed intervals of 30, 15, 10, or even six seconds. These raw scores are generally converted to a measure such as percentage of total behavior, rate per minute, proportion of deviant behavior, etc. In most cases, scores for relevant categories are summed to produce a total score in which the contribution of each response is determined by the frequency or duration with which it occurs. As an example, suppose one is using a three-category cluster to define "aggression": these are hit, yell, and tease. In a 60-minute observation, the following frequencies were recorded to yield raw aggression scores of 20, which represents 40% of all behavior observed (see Table B.1).

Here, three aggressive behaviors (i.e., variables) with different sample means and standard deviations are summed to form a composite which is heavily weighted by those behaviors with a large mean. Such variables actually lie on different metrics. In other words, TE will have a larger mean and standard deviation than HT in a sample of aggressive (or normal) children. In

traditional psychometric assessment, this situation would result in unit weighting of each variable item or score prior to combining them. But, applied behavior analysts have eschewed the use of standardization in forming composites. This procedure will be elaborated upon later. For the present, it is worth explaining the deficiencies associated with the conventional method of constructing omnibus categories.

Table B.1

	f		Percent total behavior	
Hit (HT)	= 2	Aggression = 20	4%	40%
Yell (YE)	= 3		6%	
Tease (TE)	= 15		30%	
Other negative (ON)	= 10		20%	
Appropriate (AP)	= 20		40%	
	50 total			

Generally, the magnitude of the contribution of each component behavior to a composite varies as a function of the frequency or duration with which it occurs. This may interfere with early identification and later evaluation of treatment. With respect to deviant behavior, it has been shown that the lowest frequency responses (e.g., physical negative, stealing, destruction, lying) are among the most aversive to parents (Jones, Reid, & Patterson, 1975) and to society in general. To attribute a common behavior such as "teasing" with a disproportionately higher value than lower base rate responses (e.g., "hit") is not consistent with (1) reasons for referral to treatment (Bolstad, 1974), (2) the thrust of intervention (Goodenough, 1930; Murphy, 1937; Schoggen, 1954), and (c) socially relevant criterion measures of outcome (Reid &

Patterson, 1975; Wiggins, 1973). Each of these focuses primarily, though not exclusively, on those behaviors most noxious to the social environment.

Table B.2 presents response frequencies obtained in one-hour observations of three children. In each case, the total aggression score equals 20, accounting for 40% of all behavior. Yet it is clear that child A would be described as the most aggressive, potentially the most dangerous and difficult to ignore. Child C, on the other hand, would be described as minimally aggressive by many, despite the high frequency of teasing.

Table B.2

	Child A		Child B		Child C	
Hit (HT)	8	20	2	20	0	20
Yell (YE)	8		3		1	
Tease (TE)	4		15		19	
Other negative (ON)	10		10		10	
Appropriate (AP)	20		20		20	
Total behavior	50		50		50	

Looking at the data for subject B, suppose intervention is successful in reducing HT to zero and YE to 1 in an equivalent observation period. Aggressive behavior would, at best, show a decrease from 40% to 32%, probably a non-significant improvement despite the fact that HT, the most aggressive response, was completely eliminated, and YE reduced by two-thirds. It is worth noting that an 8% decrease in aggressive behavior is the maximum possible change obtainable for the frequencies of occurrence in this example. In observation coding systems where only a portion of the on-going sequence is actually recorded, priority or hierarchical coding is performed. According to

predetermined criteria, one class of behaviors is coded instead of another simultaneously occurring response. In the above example, suppose only one behavior could be recorded per interval and that "saliency" or "aversiveness" determined which. HT would then supercede TE. If, when HT were eliminated, it produced more teasing, or simply allowed TE to be coded when previously it could not be, due to concurrence, then total aggression would remain at 40%. The ipsatizing features of the coding system itself may mask a treatment effect despite substantial decreases in low frequency, highly noxious behaviors (Jones, 1973). This fact may account, in part, for the difficulty in obtaining convergence between observation data and global ratings (Eyberg & Johnson, 1974), the latter generally showing greater improvement. In the above example, total aggressive behavior showed little change, while those responses most aversive and in fact, primarily responsible for referral, were virtually eliminated. The percentage of time during which aggressive behavior appeared remained the same; however, the profile of that behavior differed substantially following intervention. If one were interested solely in the former composite criterion measure, then only a two- or three-category coding system is required (appropriate, aggressive, other inappropriate).[1]

Presumably, the inclusion of sub-categories enables researchers to answer more specific questions about the behavioral interaction and treatment effects. Until methods are available to predict the contribution of each response category to socially relevant criterion measures, one should, at least, endeavor to control for differences in base rates by assigning weight to each

---

[1] As a general rule it is recommended that the complexity of the coding system and that of the dependent variables be commensurate. Reductions in time and expenditures of observation training and analysis would be more cost effective.

category. And then, if desirable attach weights which will produce a variable of improved construct validity. The standard scoring method described below should be used with behavioral observation data in the formation of composites.

The technique employed on the Social Learning Project at the Oregon Research Institute was adopted from traditional psychometric assessment. Often one wishes to develop an overall performance rating based upon a number of tests, each of which produces a score using a different unit of measurement. As previously stated, in cases where means and standard deviations on two instruments differ, it is appropriate to convert raw scores to standard scores for purposes of comparison. In naturalistic observation, each behavioral category can be viewed as analogous to a particular test having its own mean and standard deviation. Norms, or mean levels of frequency or duration for each category, can be obtained for a sample of normal individuals fulfilling specified criteria. Using the formula for standard z score conversion,

$$z = \frac{X - \bar{x}}{\sigma},$$

one merely substitutes the raw score of the treatment subject or group and the mean and standard deviation for the normal sample.

Returning to the example, suppose that five one-hour observations of 100 non-referred or "normal" children yielded the means and standard deviations presented in Table B.3. These children were matched on sex, age, and IQ with the cases A, B, and C. Table B.4 shows the z scores for the three illustrative subjects and the procedures used for producing category scores and a composite mean aggression score. One can readily see that the aggression score of 10.07 for child A is considerably higher than that for subject B (3.65); child C

Table B.3

Normals

	$\bar{x}$	σ
HT	.36	.48
YE	.66	.50
TE	4.05	3.81
ON	2.80	2.69
AP	42.13	17.10
Total aggressive	5.07	3.75

Table B.4

	$(X - \bar{x}n)$	=	$\div$	σn	=	z	$\bar{x}$ z aggression
Subject A							
HT	8 .36	7.64		.48		15.54	
YE	8 .66	7.34		.50		14.68	10.07
TE	4 4.05	- .50		3.81		-0.01	
ON	10 2.80	7.20		2.69		2.68	
AP	20 42.13	-22.13		17.10		-1.29	
Subject B							
HT	2 .36	1.64		.48		3.41	
YE	3 .66	2.34		.50		4.68	3.65
TE	15 4.05	10.95		3.81		2.87	
ON	10 2.80	7.20		2.69		2.68	
AP	20 42.13	-22.13		17.10		-1.29	
Subject C							
HT	0 .36	-.36		.48		-.75	
YE	1 .66	.34		.50		.68	1.28
TE	19 4.05	14.95		3.81		3.92	
ON	10 2.80	7.20		2.69		2.68	
AP	20 42.13	-22.13		17.10		-1.29	

(1.28) scored about two and a third standard deviations below child B. The transformation retains the order of severity which was lost when frequencies for each category were merely summed (see Table B.2). In effect, it translates the rates, frequencies, or durations of deviant behaviors for a particular subject or group into z scores which represent relative or comparative deviancy. A plot of obtained z scores yields a behavior profile which is readily interpretable.

Because this procedure controls for differences in base rates between responses, it is likely to yield results which favor the probability of obtaining a treatment effect. Intervention geared primarily toward the reduction of low frequency (rate) highly noxious target behaviors will, if successful, exert greater impact on a total deviant z score than on a total deviant score which is the sum of various frequencies.

Mitigating against the likelihood of obtaining a treatment effect is an increase in variance due to the transformation. The increase, which serves to reduce resultant F ratios, comes about in the following manner. A relatively uncommon response (e.g., destructiveness) will contribute relatively little to a composite score if raw frequencies for a number of categories are summed to produce a total score. Even an extremely high score for this category would have a relatively small numerical value compared to values for other higher rate behaviors (e.g., disapprove, noncompliance). The standard score for such an extreme raw value would be quite large, perhaps three or more standard deviations above the mean. When added to a standard score composite, the effect is apt to be substantial. Hence, the distribution of composite z scores is spread or widened, yielding a larger variance. So, while the transformation increases the contribution of low frequency "targeted behaviors" to a composite,

it does so at the expense of introducing additional variance which could well override the effects of unit weighting low frequency behaviors. Finally, the problem of increased variance could be recast as a problem of too little variance in the raw score (i.e., untransformed) composite.

In summary, the tendency for certain responses to covary or retain internal consistency across treatment phases lends some support to a position adhering to the conventional manner of forming observational composites. In other words, if HT decreases, so will TE. Nevertheless, the fact that internal consistency may be maintained does not psychometrically justify treating a less aversive response as more significant in a criterion measure than one of greater amplitude.