THE EFFECT OF CONTINGENCY MANAGEMENT STRATEGIES ON THE
BENDER GESTALT DIAGNOSTIC INDICATORS OF
EMOTIONALLY DISTURBED CHILDREN

DISSERTATION

Presented to the Graduate Council of the
North Texas State University in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Herbert Arthur (Chris) Ninness, B.S., M.S., Ph.D.
Denton, Texas
August, 1982
Ninness, Herbert Arthur (Chris), B.S., M.S., Ph.D., The Effect of Contingency Management Strategies on the Bender Gestalt Diagnostic Indicators of Emotionally Disturbed Children, doctoral dissertation, North Texas State University, Denton, Texas, 143 pages, 23 titles, 2 tables, 22 illustrations, appendixes, references.

Ten experimental and 10 control subjects in elementary special education were exposed to a semester of contingency management procedures for up to 6 1/2 hours per day. The experimental group was taught by teachers who were exceptionally well trained and qualified behavior analysts, while the control group was instructed by teachers with varying degrees of expertise in behavioral techniques. Both groups were given the Bender Gestalt as a pretest and posttest to determine the effect of such treatment on the diagnostic indicators of "acting out" tendencies. Furthermore, the rate of actual "acting out" was measured for all subjects by counting the number of verbal corrections resulting in placement in time out and/or warning of forthcoming time out during 20 class days of baseline and the last 20 days of treatment.

The analysis of covariance indicated a significant difference, with a p less than .05, between experimentals and controls with regard to the number of Bender Gestalt Emotional Indicators of "acting out." Also a t-test on
the pretest-posttest mean number of emotional indicators for the experimental group only showed a significant difference, at the .001. Frequencies of actual "acting out" behavior were tested for significant changes in frequency across time for each individual subject via time series analysis. Significant decreases in "acting out" frequencies were found for all the experimental subjects and half the control subjects. Thus contingency management strategies were apparently effective not only in reducing overt "acting out" behaviors but also the Bender Gestalt diagnostic indicators of such "acting out." The correlation between the mean number of such "acting out" episodes for each subject during the final 20 days of treatment and the posttest number of Bender indicators of "acting out" was .456 which was statistically significant at the .05. This correlation further substantiates the relationship between Bender Gestalt indicators of "acting out" and overt disruptive behavior as both were concurrently influenced by the contingency management procedures.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF ILLUSTRATIONS</td>
<td>v</td>
</tr>
<tr>
<td>THE EFFECT OF CONTINGENCY MANAGEMENT STRATEGIES ON THE BENDER GESTALT DIAGNOSTIC INDICATORS OF EMOTIONALLY DISTURBED CHILDREN</td>
<td></td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td></td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td></td>
</tr>
<tr>
<td>Hypotheses for Group Means</td>
<td></td>
</tr>
<tr>
<td>Research Hypotheses for Time-Series Analysis</td>
<td></td>
</tr>
<tr>
<td>Definition of Terms</td>
<td></td>
</tr>
<tr>
<td>Limitations of the Study</td>
<td></td>
</tr>
<tr>
<td>Significance of the Study</td>
<td></td>
</tr>
<tr>
<td>Review of the Literature</td>
<td>17</td>
</tr>
<tr>
<td>The Testing Controversy</td>
<td></td>
</tr>
<tr>
<td>Behavior Analysis and Personality</td>
<td></td>
</tr>
<tr>
<td>The Bender Gestalt</td>
<td></td>
</tr>
<tr>
<td>Acting Out</td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>41</td>
</tr>
<tr>
<td>Analyzing the Data</td>
<td></td>
</tr>
<tr>
<td>Subjects</td>
<td></td>
</tr>
<tr>
<td>Procedure</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>65</td>
</tr>
<tr>
<td>Discussion</td>
<td>114</td>
</tr>
<tr>
<td>General Findings</td>
<td></td>
</tr>
<tr>
<td>Conclusions</td>
<td></td>
</tr>
<tr>
<td>Recommendations</td>
<td></td>
</tr>
<tr>
<td>APPENDIXES</td>
<td>137</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>140</td>
</tr>
</tbody>
</table>

iii
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Analysis of Covariance for Experimental and Control Groups on Pretest and Posttest Bender Gestalt Emotional Indicators of &quot;Acting Out&quot; Behavior</td>
<td>69</td>
</tr>
<tr>
<td>2. Summary Table for Time-Series Intervention Analysis</td>
<td>75</td>
</tr>
</tbody>
</table>
# LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The total number of daily &quot;acting out&quot; episodes for the Experimental Group during the first 20 baseline days and the last 20 days of treatment</td>
<td>74</td>
</tr>
<tr>
<td>2.</td>
<td>The total number of daily &quot;acting out&quot; episodes for the Control Group during the first 20 baseline days and the last 20 days of treatment</td>
<td>77</td>
</tr>
<tr>
<td>3.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 1 during the first 20 baseline days and the last 20 days of treatment</td>
<td>79</td>
</tr>
<tr>
<td>4.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 2 during the first 20 baseline days and the last 20 days of treatment</td>
<td>81</td>
</tr>
<tr>
<td>5.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 3 during the first 20 baseline days and the last 20 days of treatment</td>
<td>83</td>
</tr>
<tr>
<td>6.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 4 during the first 20 baseline days and the last 20 days of treatment</td>
<td>85</td>
</tr>
<tr>
<td>7.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 5 during the first 20 baseline days and the last 20 days of treatment</td>
<td>86</td>
</tr>
<tr>
<td>8.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 6 during the first 20 baseline days and the last 20 days of treatment</td>
<td>88</td>
</tr>
<tr>
<td>9.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 7 during the first 20 baseline days and the last 20 days of treatment</td>
<td>90</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>10.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental and subject number 8 during the first 20 baseline days and the last 20 days of treatment</td>
<td>91</td>
</tr>
<tr>
<td>11.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 9 during the first 20 baseline days and the last 20 days of treatment</td>
<td>93</td>
</tr>
<tr>
<td>12.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Experimental subject number 10 during the first 20 baseline days and the last 20 days of treatment</td>
<td>94</td>
</tr>
<tr>
<td>13.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Control subject number 11 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>96</td>
</tr>
<tr>
<td>14.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Control subject number 12 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>98</td>
</tr>
<tr>
<td>15.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Control subject number 13 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>99</td>
</tr>
<tr>
<td>16.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Control subject number 14 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>101</td>
</tr>
<tr>
<td>17.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Control subject number 15 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>102</td>
</tr>
<tr>
<td>18.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Control subject number 16 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>104</td>
</tr>
<tr>
<td>19.</td>
<td>The total number of daily &quot;acting out&quot; episodes for Control subject number 19 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>105</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>20. The total number of daily &quot;acting out&quot; episodes for Control subject number 20 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>21. The total number of daily &quot;acting out&quot; episodes for Control subject number 21 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>22. The total number of daily &quot;acting out&quot; episodes for Control subject number 22 during the first 20 days of baseline and the last 20 days of treatment</td>
<td>110</td>
<td></td>
</tr>
</tbody>
</table>
THE EFFECT OF CONTINGENCY MANAGEMENT STRATEGIES ON THE
BENDER GESTALT DIAGNOSTIC INDICATORS OF
EMOTIONALLY DISTURBED CHILDREN

Emotional disturbance in the public school system pertains to
students whose emotional condition is psychologically or psychiatrically determined to be such that they cannot be adequately and safely educated in the regular classes of public schools without the provision of special services. (Texas Education Agency, 1979, p. 22)

The key words relating to adequately and safely convey particular concern. Such children are defined divergently throughout the literature as suffering from differing disorders, but within the public school these words often pertain to the need to attend to children whose behavior is disruptive, bizarre, extremely inappropriate, and basically not capable of being controlled in the typical classroom setting "adequately" and "safely."

There exist within the discipline of school psychology many different and diversified approaches to the study of such children. Emotionally disturbed children have traditionally been educated, counseled, and subjected to
therapy according to variations of the classic Freudian psychoanalytic theory (Bettleheim, 1955; Freud, 1937; Reinert, 1978). The basis of such intervention emphasizes a determination of the underlying "meaning" of disturbed behaviors which are ostensibly a function of unresolved, unconscious, regressed development of primary and secondary processes relating to the critical life periods (Corsini, 1977). According to a review conducted by Pelton (1981), the educational models which have evolved from the traditional analytic approach frequently propose a rather permissive nonstructured academic atmosphere for the emotionally disturbed child. It is assumed that exploration of individual needs is more critical than the teaching of academic skills (Bettleheim, 1955).

There are a multitude of variations on the basic psychodynamic theme; however, there is a pervasive fundamental theory relating to a need for "insight" into the dynamics of the child's problem. In all such variations, the subconscious problem is the primary concern. The manner in which the various analytical and neoanalytic schools attempt to resolve such problems is highly diversified and often tangential to the basic tenets laid down by Freud.

Free play is encouraged by Jungians who hope to alleviate the stresses associated with regression. They
attempt to analyze subconscious symbolic meanings in the child's fantasies (Maduro & Wheelwright, 1977). Drichers (1964) extends the Adlerian model into joint class discussion where insight arises as a function of feedback from peers. Axline (1947) is a particularly good example of tangential psychodynamic theorizing. She attempts to have the child grasp "insight into his underlying problem" by permissive interactions with his instructor/therapist. She even specifies that academic instruction is to be deemphasized in favor of attending to the child's underlying needs. The format for such insight acquisition is group or individual psychodynamic play. In much of the same vein, Rogers (1942, 1952) has proposed unconditioned positive regard and an emphatic understanding of the internal frame of reference. Where this occurs, a process of change is in motion. During the process of change, the student/client becomes increasingly aware of his true feelings and experiences and his true self-concept. Teachers are told to allow, as with Axline, freedom of expression. They are told to reorganize basic social structures to fit the idiosyncrasies of the child (Pelton, 1981). Defiance is to be accepted and punitive actions are reduced or eliminated while setting the stage for internal discovery. Gestalt psychology has not missed its opportunity to comment on the enigma of
emotional disturbance. Here again we hear of teaching self-awareness, promoting self-actualization, and structuring the child's environment in such a way that he gains a sense of control of his environment (Pears, 1959). Peak experiences (euphyschia) are emphasized by existential psychologists who address many of the same notions expressed by the Gestaltins. Paramount within this analysis is the need to develop one's sense of creativity and the sense of self (Corsini, 1977; Pelton, 1981).

Throughout the realm of the psychodynamic educational position runs the common thread of combining a therapeutic milieu with an educational process (Rothman, 1974). In contrast to previous analytic and derivatives of analytic models which have traditionally been popular in the school systems, the behavioral model emphasizes learning as the primary mode of intervention. The previously mentioned medically oriented models either directly or by implication have attempted to resolve underlying maladies by "curing" the disturbance or by providing insight into an evasive, unconscious source of disturbance. For the behavior analyst, the overt inappropriate behavior is the problem. And the problem behavior is dealt with directly. Emotional disturbance is understood to be a function of either unlearned skills, inappropriate to the classroom protocol, or learned avoidance behaviors. If
such inappropriate behavior is learned it can be unlearned. If the problem is one of lack of skills, social or academic, they can be taught (Whaley & Mallott, 1968). O'Leary and O'Leary (1972) provide a warehouse of behavioral strategies that can be used in the classroom to teach "disturbed children appropriate skills." Rinehart (1978) has reviewed the current literature relating to classroom management and notes that a new era of remediation and therapy has begun with the advent of behavioral technology.

Procedures adopted by Engleman, Carnine, and Johnson (1978) have allowed the teacher to structure the class in such a way that all students can respond and be reinforced in unison. Other procedures, developed by Homme (1971) and Premach (1959), provide techniques for improving the quality and duration of on task behavior. Long and Williams (1973) have designed classroom contingencies which provide consequences for group behavior, thus organizing a kind of peer pressure for proper classroom conduct. The literature is replete with examples of applications of behavioral methodology in the classroom (Salzer & Mayer, 1972; Scott & Bushell, 1974; Zimmerman, Zimmerman, Rider, Smith, & Dinn, 1971).

**Statement of the Problem**

The problem tested in the following study is the degree to which behavior therapy and classroom management
is capable of effecting not only particular target behaviors but also the more subtle and broad range personality factors as measured by a popular index of emotional disturbance. In particular, this research endeavored to analyze the effects of a classroom contingency management program on the rate of disruptive ("acting out") behaviors of emotionally disturbed children in conjunction with the changes in Bender Gestalt protocols.

**Purpose of the Study**

The purpose of the study is to determine the relative effectiveness of a classroom contingency management procedure on children who perform high rates of inappropriate, disruptive, and bizarre behaviors in the classroom. These children were designated as showing strong indices of emotional disturbance on the Bender Gestalt, a test typically used by the more traditionally oriented school psychologists. The behavioral measure of rate of inappropriate, disruptive behaviors was also taken throughout the semester. The degree to which a behavioral contingency management program is capable of influencing frequencies of disruption, together with emotional indices of disturbance on the Bender Gestalt, provide evidence concerning the efficacy (or lack of efficacy) for procedures, popularly known as "behavior modification," to influence outcomes often referred to as "emotional indicators" of a disturbed "personality."
A more global purpose of this research is to address the common discord between the behaviorally oriented psychologist and the traditional psychologist. The degree to which measures of behavior, representative of both academic biases, are influenced by a behavioral intervention has serious ramifications for psychologists in all of the "applied" areas. The extent to which no significant effect is determined is perhaps equally critical. If behavioral strategies, in fact, influence emotional indices on the Bender, then perhaps psychologists of behavioral orientation should incorporate a broader scheme of dependent variables into their outcome measures. Also, such outcomes are likely to provide the traditional/analytic psychologist with meaningful information concerning the extent to which current environmental influences provide a generalizable effect throughout the individual's whole "personality." If no evidence for change is seen in the Bender protocols while rates of disruptive behavior are retarded, this would add more certainty to the general behavioral disposition that such "personality measures" are basically irrelevant. Such an outcome will no doubt tell the traditional psychologist that behavioral tactics do not effect the underlying factors of emotional disturbance. If neither reduced rate of disruptive classroom behaviors nor changes in Bender protocols are affected by the contingency management tactics, then this study will have produced no evidence
worthy of consideration for existing theories, but perhaps questions of methodology may be further analyzed. Details of the extensive ramifications for this research are elaborated throughout the course of the following pages.

**Hypotheses for Group Means**

The major problem to be advanced by this research is the degree to which a reduction in "acting out" (disruptive) behavior in the classroom as a function of reinforcing appropriate classroom behavior and peer interaction ultimately effects the number of Emotional Indicators on the Bender Gestalt test. The following hypotheses were specified previous to data collection.

1. The null hypothesis (H₀) for the analysis of covariance states that there will be no posttest differences in the means of Emotional Indicators on the Bender Gestalt between the experimental contingency management group and the standard special education control group.

2. The research hypothesis (H₁) for the analysis of covariance states that there is a difference in the posttest means of Emotional Indicators between the two groups.

3. The null hypothesis (H₀) for the t test on the experimental group states there are no differences in pre- and postintervention means of Emotional Indicators on the Bender Gestalt.

4. The research hypothesis (H₁) for the t test on the experimental group states there will be a difference
between pre- and posttest means of Emotional Indicators on the Bender Gestalt.

Research Hypotheses for Time-Series Analysis

1. It is also hypothesized that there will be a significant decrease in the frequencies of "acting out" behaviors, over the period of the entire semester, for the contingency management group as a whole.

2. There will correspondingly be a series of significant differences in the "acting out" frequencies of each individual subject in the contingency management experimental group. (Note: Both individual and group frequencies may be analyzed via time series analysis.)

3. It is anticipated that there will be no significant differences in the frequencies of "acting out" behaviors of the control group as measured over time.

4. There will be a corresponding lack of significance in the "acting out" frequencies of individual members of the control group.

5. In a more general sense, it is hypothesized that a decrease in "acting out" behaviors, which occur as a function of differentially reinforcing increased positive interactions among students and between staff and students, will, because of common contingencies, eventually be reflected in a significant decrease in the number of indicators of emotional disturbance on the Bender Gestalt.
Definition of Terms

Behavioral Technology. Models of conditioning techniques which are based on procedures which emphasize the shaping of behavior by the judicious use of consequences.

Classical Desensitization. A behavior therapy technique which utilizes the processes of associating relaxation with a previously feared stimulus.

Classroom Management. Any system which imposes group or individual environmental contingencies on the students in a class. Such systems usually exchange tokens or tangible reinforcers for increasing amounts of appropriate classroom behavior or provide for varying degrees of punishment for inappropriate behavior.

Contingency. The extent to which the occurrence of a given behavior is dependent on a specified environmental event.

Contingency Management. A behaviorally oriented classroom program which emphasizes the degree to which appropriate social and academic behavior allows students access to social and tangible reinforcers. Inappropriate classroom behaviors are typically ignored or result in temporary isolation from class.

Defiance. A client's resistance to psychological intervention in such a way as to connote open or covert hostility.
Existential Psychology. A derivative of existential philosophy which emphasizes subjective experience and the irreducible uniqueness of the individual. This position regards human experience as being undescrivable in idealistic or scientific terms and stresses the need for individual freedom and responsibility.

Fear Survey Schedule. A behavioral assessment instrument which ranks various behaviors and events in degree of anxiety provocation.

Neoanalytic. Revised versions of psychoanalytic theory which typically emphasize the social rather than sexual factors which underlie motivational processes.

On Task Behavior. Behavior which is operationally defined as having students actively perform academic behavior in such a way as to be verifiable by an impartial observer. For example, students must be seated in their seats writing with appropriate implements for a specified period of time.

Psychoanalytic Theory. A systematic approach to the understanding and cure of "mental" disorders, developed (Originally) by Sigmund Freud. This view holds that the "roots" of human behavior are to be found in unconscious motivation and conflict (Wenrich, 1970, p. 83).
**Psychodynamic.** Refers to psychoanalytic and other "depth" psychologies. The term also implies a major concern with the study of emotional and motivational processes (Wenrich, 1970, p. 83).

**Stress.** Any environmental or internal form of stimulation which produces a direct or indirect deliterious effect and typically results in an attempt on the part of that organism to escape or avoid such stimulation.

**Successive Approximation.** A method of reinforcement whereby the organism receives reinforcement only on those occasions during which he performs behavior which more closely approximates a given target behavior.

**Regression.** Have reached a certain stage of development, a person may retreat to an earlier level because of fear (Hall, 1954, p. 95).

**Unconditional Positive Regard.** A therapeutic posture, originally formulated by Carl Rogers, which emphasizes the necessity of the therapist's maintenance of a caring and positive attitude toward the client independent of any aversive dispositions of the client.

**Variable Interval Schedule (VI) of Reinforcement.** A reinforcement schedule that specifies that the first
response occurring after a given period of time has elapsed, since the last reinforced response, will be reinforced. The time intervals between such reinforcement fluctuate around a specified mean interval.

Limitations of the Study

One of the obvious limitations of the following study involves the relatively small group size. Use of such small Ns not only reduces the power of one of the proposed statistical procedures but also reduces the degree to which results of this particular investigation can be generalized to populations outside the sample. However, it is important to note that if and when statistical significance is obtained with small Ns, the degree of effect of the treatment must be assumed as indicative of an intense effect. Further, a small N is required in this context in order to provide all subjects with the individualization required to produce an effect. It is only by increasing the teacher's attention to the students' appropriate behavior that such behavior can increase in frequency and thus, it is hoped, produce a more appropriate student repertoire. This has required a small student-to-teacher ratio.

A more complicated methodological limitation involved the possible confounding of treatment and measurement with regard to one of the criteria for effect of treatment.
The most critical aspect of treatment involved the social and tangible reinforcement of students' increasing tendencies toward appropriate classroom behavior and peer interaction. In most cases this involved the teacher or psychologist verbally commending the giving tokens to students as they gradually assumed behaviors which were seen as appropriate to the classroom context. This also involved teaching the children rules regarding social interaction with peers and authority figures. It is this process which was anticipated to produce an eventual change in the child's repertoire such that he is less inclined to "act out" in the classroom and less inclined to show indices of emotional disturbance on the Bender Gestalt.

However, it will be noted that one criterion for the effect of this treatment is a reduced frequency in the "acting out" behavior as designated by placement in time out or verbal warning. But placement in time out and verbal warnings constitute treatment in themselves. As such, the criterion for measuring the effect of reinforcement of appropriate behavior is partially confounded with another treatment, the use of any of the corrective procedures (time out, warning). One could perhaps argue that any forthcoming reductions in the frequency of "acting out" are not only a function of reinforcing appropriate
behavior, but are at least partially a function of the use of such corrective measures.

This may be an empirical question, one which can best be answered by analyzing the baseline frequencies and baseline contingencies. Baseline contingencies for experimentalss as well as controls do involve the consistent use and measurement of the previously mentioned corrective actions. Therefore, if they in themselves constitute a treatment effect in terms of reduced frequencies of "acting out," this should be apparent by visually inspecting the frequencies of the exceptionally and necessarily elongated baseline period (20 days). To the extent that any gradual reductions in such "acting out" frequencies are discernible, on baseline, this militates against a reinforcement only effect. However, even if such indications are apparent on baseline, this will be an important source of interaction to be later analyzed and discussed.

If these corrective procedures (i.e., time out) are not found to demonstrate reduced frequencies during baseline, this would hardly be a surprising outcome. The use of time out, and warnings for forthcoming time out, are typically effective only when dense reinforcement schedules are in effect. These dense reinforcement schedules will not become a prominent part of the
contingency management program until such time as treatment begins. It will be recalled that time out literally means time out from reinforcement.

Significance of the Study

The extent to which the two dependent variables (frequency of "acting out" and number of Emotional Indicators on the Bender Gestalt test) are commonly affected by the independent variable (classroom contingency management) provides evidence concerning the wide range effects of behavior therapy in conjunction with continuing validity data for use of the Bender Gestalt as an assessment instrument of emotional disturbance. There is a continually growing need for assessment techniques which are not only sensitive to the extent of individual severity of emotional disturbance, but also sensitive to the degree to which intervention technique ameliorates these indices of disturbance. Currently, there are no widely accepted personality tests which are seen by behavior therapists as being sensitive to general improvements in the individual's social repertoire. If the Bender Gestalt proves to be consistent with concurrent changes in frequency of "acting out" behaviors, this would be strong evidence for continued use in a behavior therapy armamentarium. It would provide all psychologists with further information concerning ways to determine the
pervasiveness of their therapeutic strategies. And, it would, in the long run, contribute to the benefit of the client population.

**Review of Literature**

**The Testing Controversy**

Within the realm of school psychology there are many competing influences which seek to bring about effective psychological improvement on the emotionally disturbed student. These diversified forces fall basically under the heading of traditional psychoanalytic/neoanalytic, humanistic, and the behavioral approach. These systems and their adherents have styles and methodologies which are at times bitterly opposed but occasionally converge in an overlap of theory and practice. However, the most conspicuous point of contention among these positions revolves around the implementation and necessity of testing and, most particularly, within the realm of what has been called personality testing and projective techniques.

Testing has been a required element within school psychology because it has become necessary to categorize and place students who deviate academically and behaviorally. Testing has, therefore, been a tool for the diagnostician, psychometrician, and psychologist to assist in the placement of the deviant student within the school system. This has been a particularly popular strategy
where emotionally disturbed children are concerned. The use of "personality tests" has supplied the medium through which justification for and admission to special education classes have been determined.

Though such emotionally disturbed children are legally required to be placed in the "least restrictive environment" (Public Law 94-124), the probability of one being dismissed from special education placement is typically not based on the criterion which was originally used to admit the student, namely the personality test. Rather, dismissal from special education placement, when it occurs, is most frequently predicated on Admission, Review, Dismissal committees' (composed of parents, teachers, counselors, therapists, and psychologists, and principals) decisions which reflect a change in the classroom behavior of the child, not on test performance. In other words, though personality tests are used as input, they are seldom, if ever, used as a criterion for output with regard to dismissal from a special education program. Such tests are not used as criteria for dismissal because they are typically not seen as sensitive to changes in behavior as a function of therapy conducted in the schools. Though the emotional indicators depicted on a projective technique often do correlate with undesirable behavior (Caudle, 1981), concern becomes warranted when these tests
are used as an explanation of the deviant behavior (Koppitz, 1968). Tests which are purported to diagnose behavior but are not necessarily sensitive to changes in behavior as a function of intervention may be seen as somewhat teleological. The "personality" tests are seemingly sensitive to pathology but not to the variables for which the pathological behavior is a function or even the remediation of such pathological behavior. Thus, these diagnostic tools provide explanations which seemingly add information concerning the child's pathology but in fact only change the way in which he is described. However, because these diagnostic interpretations seemingly provide a sufficient explanation of a given phenomenon, they in fact postpone a more functional analysis which might be aimed at determining what variables are responsible for both the emotional indices of pathology and the behavioral manifestations of it.

Criticisms such as these have forestalled behaviorally oriented psychologists from relying on or even being very much interested in the use of so called "personality tests." On the other hand, behavior therapy has typically been seen by much of the traditional psychological community as a series of "quick and dirty" techniques which may change a person's overt behavior pattern, for the present, but actually have little if any effect on the underlying
problems that may be at the core of an individual's emotional turmoil (Rimm & Sommervill, 1977). These emotional indices as measured by several of the more subtle cognitive or dynamically oriented personality tests are said to be not easily altered or improved by strategies which are aimed at merely "modifying" a person's overt behavior. It has often been stated that the complexity of human behavior cannot be changed by procedures which deal only with such superficial behavior. A typical comment to this effect is made by Sunberg and Tyler (1962). They state, in discussing behavioral procedures, in their classic text Clinical Psychology that "Experienced psychotherapists are so much aware of the complexity of human personality that systems and methods like Wolpe's inevitably impress them as oversimplifications" (p. 351). They also point out that

the therapist who chooses to work according to the procedures we have been discussing [behavior therapy] finds little use for the customary assessment procedures, with the exception of interviewing. He [the behavior therapist] does not ordinarily think in terms that are used by psychological diagnosticians.

(p. 347)

This is not entirely an unfair statement. It has usually been held by behaviorists that direct observation
of behavior is more indicative of the real world than inferences drawn from personality tests; even when such tests have been devised with sophisticated factorial statistical procedures (Bijou, Peterson, & Ault, 1968; O'Leary, 1972). When intervention is performed by a behavior therapist it is usually in an attempt to change the frequency of a specific behavior or to shape a new response from the elements of behavior already existing within an individual's repertorie. This, of course, presupposes a target behavior and any assessment of the effect of such intervention will be predicated on reaching that target behavior. Other responses in the person's repertoire may or may not be effected by this intervention. The goal of the behavior therapist is aimed primarily at this target behavior independent of any ancillary response variables effected by treatment. Although this makes for a rather straightforward interpretation of the degree of effect, it says nothing about the individual's propensity to perform other behaviors not associated with the treatment.

Personality tests and inventories are much less likely to be structured around such idiosyncratic response variables. Personality tests in general, according to Anastasi (1968), are "instruments for the measurement of emotional, motivational, interpersonal, and attitudinal characteristics, as distinguished from abilities" (p. 437). Further,
unlike behavioral assessment strategies which were born out of the rigors of the experimental laboratory, personality tests have originated in clinical settings. Again, according to Anastasi (1968), "the amount of experimental verification to which they have subsequently been subjected varies tremendously from one theoretical system to another" (p. 583).

While behavior therapists have traditionally been somewhat obsessed with measures of reliability, many of the more clinically oriented testing devices have not considered this to be as an important a consideration. This is particularly true within the realm of projective techniques. In this regard, Anastasi (1968) points out that few adequate studies have been conducted on the scorer reliability of projective tests. Several investigators have revealed marked divergencies in the interpretations given by reasonably well-qualified test users. A fundamental ambiguity in such tests results from the unknown contribution of the interpreter's skill. Neither high nor low scorer realibility can be directly generalized to other scorer's differing appreciably from those utilized in the particular investigation. (p. 579)

Thus, it appears somewhat obvious as to why research oriented behavior therapists have traditionally avoided
the ambiguities of traditional personality tests in general and projective techniques in particular. They have willingly sacrificed the broad based assessment of the "whole" integrative personality, which is somewhat elusive, for the more poignant but measurable assessment of particular behaviors as effected by treatment applications.

Following Skinner's (1938) dictum they have most often chosen rate as their primary criterion of change. Although measures of response intensity, duration, and magnitude have likewise been investigated, the acceleration and deceleration of response rate have proven the most amenable to the measurement of behavior change. Rate as indexed in terms of response per unit of time has proven to be an accessible medium across species and treatments (Millenson & Leslie, 1979). Cumulative records taken from pigeons, rats, monkeys, and men have produced remarkably similar records when these organisms have been responding on the same schedule.

Rate or frequency of response per time unit is typically measured within the confines of particular response classes. These response classes or operants are defined functionally as all behaviors that effect a specified amount of environmental change, e.g., door opening responses. Frequently, operants are defined topographically as organismic movements that fall within the physical capacity of that species, e.g., hand raising responses.
Both of these definitions are easily defined operationally and easily observed. Most importantly, the reliability of the frequency of such operants can be easily quantified as the percent agreement between two concurrent observers.

But the degree of scientific reliability in behavior analysis has seldom come into question. The criticism from competing schools has more frequently been in regard to what changes in mean response strength mean when defining the "whole person." The question has been how, in the context of the elementary school, does a change in the rate of "talking out," "out of seat behavior," or "aggressive behavior," as defined operantly, affect the more global emotional, motivational, interpersonal, and attitudinal characteristics (Anastasi, 1968), as measured by traditional personality tests?

This can be defined as a question of generalization (in the broadest sense of the word). A behavioral definition of stimulus generalization relates to the tendency of an organism to respond to stimuli which approximate the training stimulus within the confines of a previously reinforced response class (Whaley & Mollott, 1968). But in a less structured clinical sense, the question may be one of to what degree do the behaviors reinforced in one environmental context tend to be emitted in other environments? More practically, what are the chances of a child
who is shaped by successive approximations into appropriate classroom behavior behaving properly on the school bus or playground? More poignantly, what are the chances that behavior shaped in the classroom will be internalized by the student such that he more closely approximates "emotional, motivational, interpersonal, and attitudinal characteristics" defined as normal by various personality tests? This is particularly in question when the student who exemplifies inappropriate, disruptive, and bizarre behavior in the classroom likewise displays emotional indicators of disturbance on the personality test. Changing the classroom behavior by way of behavioral contingencies still leaves the question as to what extent, if any, the student has had his "personality" altered by those procedures.

**Behavior Analysis and Personality**

Despite these seeming inconsistencies between behavior analysis and personality evaluation there have been a small number of studies which have attempted to determine the impact of behavior therapy on personality. A study conducted by Tanner (1971) found that use of classical desensitization was effective in not only eliminating 10 compulsive behaviors of a 24-year-old female student, but also affected the girl's profile of elevations on the Minnesota Multiphasic Personality Inventory (MMPI). This corresponded with a similar reduction in the Fear Survey
Schedule. The direction and extent of improvement indicated a broad range overall adjustment.

Another small study was conducted by Johnson (1980) in which he found a "proclivity" of support for the notion that personality type of three developmentally disabled adolescents, as measured by the Myers-Briggs Type Indicators, became more like the personalities of their counselors after extensive use of a token economy system. Though Reynolds (1981) performed no explicit behavioral intervention in the classroom, he did find a relationship between self-esteem, as measured by the Coopersmith Self-Esteem Inventory, and classroom behavior, as measured by the Classroom Behavior rating scale. Reynolds speculates that this phenomenon is due to differential reinforcement of appropriate behaviors in a large number of individuals and the use of punishment and negative reinforcement with others. "Self-esteem" is thus shaped by the verbal commentary of the classroom teacher. If this is, in fact, true, then the classroom should prove an excellent milieu for intervention with those individuals suffering from low self-esteem.

Another study which did provide personality outcome measures as a function of behavioral intervention strategies was conducted by Jesness (1975). In an attempt to determine whether behavioral or transactional analysis
strategies were most effective with delinquent youth, subjects were given tests for personality change in the form for the Jesness Inventory and tested for behavioral change using the Jesness Behavior Checklist. Not surprisingly, it was found that those delinquents exposed to behavioral methodology showed the largest change on the behavioral checklist while those who were treated with transactional analysis showed the greatest gains in the personality indices. Kennedy (1978) found that teachers using the Devereux Elementary School Behavior Rating Scale to evaluate self-concept related closely to students' own personal evaluation of self as determined by the Self-Observation Scale. Further, the teachers' prediction of tendencies toward disruptive classroom behavior on the Devereux corresponded to the students' self-concept to the teachers' prediction of the children's self-concept. Kennedy concludes that teachers "must attend to the students' anxiety in the classroom to reduce disruptive behavior" (p. 73). This anxiety is assumed to be at least partially related to learning material difficulty and it is, therefore, recommended by Kennedy that teachers pay particular attention to this variable.

In an article curiously entitled "Self-Concept Changes Following Behavior Modification," Morrow (1975) described a modified version of Bills' "Index of Adjustment and
Values" (Bills, 1961), used to measure self-concept, ideal-self, and self versus ideal-self discrepancy in two fourth grade classes, one fifth grade class, and two sixth grade classes of inner city schools. Subsequently, behavior modification strategies were incorporated to see what effect these strategies had on the Index of Adjustment and Values. Unfortunately, only five subjects in all the classes were selected by the teachers for this behavior intervention and the results could not be deemed conclusive. Another study was presented by this author in conjunction with the former. In that study a social worker conducted classes in behavior modification with parents who requested guidelines in child rearing. These parents were pre- and posttested using the same Index of Adjustments and Values. No significant gains were found between experimental and control groups. The author speculates that the treatment was "not sufficiently potent and salient" enough to produce efforts in self-esteem. He also questions whether his assessment tool was sensitive to the effects of his treatment. The title, specifying "self-concept changes," seems somewhat incongruous with the outcome measures or lack of same.

Lack of significance was also found in a study by Dixon (1978) in which 82 undergraduates were pre- and posttested using Rotters Internal-External Locus of
Control Ideology Form, the Rokeach Dogmatism Scale, the
Pupil Control Ideology Form, and three factors on the IGPH.
Subjects were exposed to "an extensive classroom behavior
modification laboratory." Again, as in the previous study,
no statistical significance was found. (Perhaps this is
another case of insensitive outcome measures.)

An excellent review of assessment in behavior modifi-
cation strategies by Kanfer (1972) points out the "lack of
standardized and widely used procedures." The study indi-
cates that most of behavioral psychologists are inclined
to use life history questionnaires, activity schedules,
behavioral observation, checklists, and some biological
assessment. He further pointed out that behaviorally
oriented therapists have shied away from the traditional
psychological forms of assessments, i.e., personality
tests. Basically the same assertions are put forth by
Goldfried (1968).

The reader of current literature in this area of
personality testing and behavior analysis is usually left
with the notion that these two specializations within
psychology have such deeply ingrained biases against each
other that their incorporation into a useful intervention/
outcome system seems highly unlikely and unadvisable.
This is true not only of academic prejudices but also
because the aficionados of these respective specializations
are usually just not interested in each other's description of behavior.

Behaviorists do not care about such phenomena as confabulation and the psychodynamically oriented are usually not interested in response rate as an index of behavior change. But perhaps there are, as of yet, untried possibilities.

Recent research by Caudle (1981) has unearthed some exciting innovations with regard to at least one measure of emotional disturbance via a well worn clinical/personality, perceptual-motor, developmental assessment tool. The Bender Visual Motor Gestalt test has surfaced as a surprising correlative index of disruptive behavior in the classroom. Before elaborating on the specifics of this research, it would seem advisable to revisit the basic procedures and recent literature related to this test.

**The Bender Gestalt**

The Bender Gestalt test is a rather straightforward testing procedure wherein the subject is asked to replicate nine different 4 X 6 inch cards which are placed before him, one after another. He is given a piece of 8½ X 11 inch blank typing paper and a No. 2 lead pencil and told to copy the designs on these cards. Subjects are typically given as much time and paper as desired but must use no more than one sheet of paper and time rarely lasts longer than about 15 minutes. The
average duration of the test runs at about 6 minutes, 20 seconds (Ackerman, Peters, & Dykman, 1971; Dykman, Peters, & Ackerman, 1973). Currently, the Orthopsychiatric Association (Bender, 1946) regulates the standards and publication of these test cards. Replicas of the specific card designs, at 75% of original size, are depicted in Figure 1.

These designs were originated in an experiment conducted by Wertheimer (1923) in his early Gestalt investigations concerning integrative processes. Wertheimer was studying organic brain disorders as well as the intellectual efficiencies of normals. Since even Bender (1967) maintains that Bender Gestalt retests of the same individual will show great fluctuations in the specific indicators of emotionality, it becomes very important to document reliability studies related to this test. A number of such studies have been conducted and are well documented by Koppitz (1973). A few of the more salient studies will be described. Isaac (1973) found no statistically significant differences on test-retest score with first grade children tested on the same day. Goff and Parker (1969) found significant correlations at .83 to .90 for students in kindergarten to fourth grade. In this study, there was a 2-week interval between tests. Rychman, Rentfrow, Fargo, and McCartin (1972) tested lower
Fig. 1.—Duplications of the Bender Gestalt cards at 75 percent of the standard size.
middle-class children in the second, fourth, and sixth grades. On retests performed 1 week later correlations ranged from .53 to .76. The above reliabilities best exemplify populations reminiscent of those in the forthcoming study and appear satisfactory for research purposes. It is important to note that such factors as increased motivation (Isaac, 1973), familiarization with the test (Zach & Kaufman, 1972), and even training in perceptual-motor tasks (McQuin, 1967) did not seem to significantly influence test-retest results.

Though typically associated with perceptual-motor performance, there are a number of highly diversified and multifaceted ways in which the Bender Visual Motor Gestalt Test (Bender, 1938, 1946) has been administered and interpreted with both clinical and school populations (Caudle, 1981). Diagnosis of such phenomena as organic brain disorders (Hartlage, 1970; McConnel, 1967; Stavrianos, 1970), learning dysfunctions (Ackerman et al., 1971; Paul, 1971), mental retardation (Simensen, 1974), and screening for readiness for school attendance (Koppitz, 1971, 1973, 1975; NorFleet, 1973) as well as school achievement (Dibner & Korn, 1969; Koppitz, 1971, 1973, 1975) have been accomplished. The Bender has been used as a projective technique with apparent psychoanalytic overtones by Lerner (1972), and as an index of
the effects of Electroconvulsive Shock (Erwin & Hampe, 1966). Koppitz (1973) notes that Fischer (1973) "does not analyze Bender records at all; instead she 'goofs around' with the Bender Test" (p. 9). Apparently a vast number of possibilities exist. Koppitz (1973) indicates the Bender ranks third in popularity with clinicians as the test used most frequently with the majority of their clients.

In an excellent review of the literature pertaining to the various uses of the Bender-Gestalt, Caudle (1981) points out that Hutt (1950) was one of the originators of using the Bender-Gestalt as an index of emotional disturbance. In the early work, 26 different indices of emotional disturbance were recognized. These included such components as design organization, size, distortions, and position of the individual relative to the test paper. Later studies (Byrd, 1956) using factor analysis indicated characteristic emotional disturbance determinants associated with notations (the tendency to draw the design on a different angle from which it is observed), incomplete closure (failure to close the design), reorder of presentation sequency, and curvature change (changing the direction and angulation of curves in the designs). Clawson (1959) initiated the differential examination of clinical population by testing disruptive, undisciplined children and
withdrawn children. It was found that five indicators correlated with these mutually exclusive behaviors. Expansion, horizontal positioning, and uneven figures were most reminiscent of those who were most inclined toward disruption while small designs and compressed organization were seen most often in children diagnosed as withdrawn. A concurrent study by Clawson found an additional 13 indicators of differences between normal public school children and those in a child guidance center.

Subsequent studies utilizing differential diagnosis of the disturbed child have included Goldberg's (1957) description of protocols performed by schizophrenic versus normal children, Eber's (1958) study of well versus poorly adjusted retarded children, Zolik's (1958) delinquent, nondelinquent investigation, and Corotto and Curnutt's (1960) examination of adolescents who were depicted as either aggressive or inclined toward flight when given a frustrating circumstance.

Cross cultural studies with children using the emotional indicators on the Bender have been conducted by Clifford (1977) and Lifshitz (1978). Clifford found increased numbers of emotional indicators in children diagnosed as disturbed and Lifshitz found correlations in emotional indices pertaining to aggression and withdrawal with teacher ratings describing the same phenomena.
Though there are a number of different scoring procedures popularly used with the Bender, most of these are capable of differentiating divergent groups (Koppitz, 1973). One of the more popular but complete scoring systems was developed by Pascal and Suttell (1951). However, even this relatively elaborate and adult oriented scoring system was well adapted to discriminating between well adjusted and poorly adjusted third-grade boys (Fromm, 1966).

**Acting Out**

Emotional indicators relating to what has historically been referred to as "acting out" behaviors have been well correlated with specific determinants on the Bender. "Acting out" is a somewhat overused term which relates to reducing neurotic anxiety by relieving the stress which the id has exerted on the ego. This is accomplished by performing an impulsive act. However, such an outward manifestation of internal pressure is said to escalate reality anxiety because the impulsive act evokes a threatening reaction from the environment (Hall, 1954). "Acting out" may for more empirical purposes be understood to be somewhat synonymous with disruptive behavior. The literature has, however, been somewhat evasive concerning this issue.
Independent of the term's origin, "acting out" has been well predicated with the Bender. Aggressive, impulsive, disruptive children appear to be inclined toward producing specific indicators on their protocols. Brannigan, Barone, and Margolis (1978) demonstrated this with 60 elementary school children. They found significantly more of such emotional indicators of impulsivity as compared with a control group who behaved normally. Likewise, Mogin (1968), Handler and McIntosh (1971) found a relationship between impulsive, disruptive behaviors as measured by teacher ratings and the determinants on the Bender protocols. Another study by Brannigan and Benowitz (1975) arrives at basically the same conclusion when comparing Bender protocols with ratings by house parents of "antisocial acting out tendencies." The list of such comparative relationships between "acting out," aggression, impulsivity, and general disruptive behaviors and Bender's indicators of such phenomena goes on and on (DeCato, 1976; Koppitz, 1964, 1971; Lerner, 1972; Toler, 1980). These and many other studies relating to the emotional indices of "acting out" are well documented by Caudle (1981).

Caudle, in his search for Bender indicators of "acting out" behavior, isolated six such determinants that had not been specified by Koppitz (1963, 1973). These
were (a) a progressive increase in size from Figure 1 to Figure 8 as documented originally by Hutt (1977) and Clawson (1959), (b) shape angles on Figure 6 specified by Brown (1965) and Hutt (1977), (c) collision of designs as elaborated by Brown (1965), Hutt (1977), Lerner (1972), and Tolor (1968), (d) dashes substituted for dots in any of Figures 1, 3, or 5 (Brown, 1965; Hutt, 1977; Brannigan et al., 1978; Zolik, 1958), (e) circles substituting for dots in any of Figures 1, 3, or 5 (Brown, 1965; Brannigan et al., 1978; Zolik, 1958), and (f) uneven figure size (Brannigan & Benowitz, 1975; Clawson, 1959). In addition, Caudle noted six of Koppitz's (1973) emotional indicators as being reminiscent of "acting out" behaviors.

Since such overwhelming evidence for a relationship between "acting out" behaviors as measured by teachers, houseparents, counselors, and others, and indices of emotionality on Bender protocols existed, Caudle noted the need for a study which would replace the rating scales previously used with actual frequencies of aggressive, impulsive, disruptive behaviors. He hypothesized that these six plus six of Koppitz's determinants would accurately predict the frequency of acting of juveniles placed in an alternative education program. His criterion for specifying an occasion of "acting out" behavior was based on the necessity of a student being
placed in "time out" ("time out" is a disciplinary procedure used by many behaviorally oriented programs, which places a disruptive student in isolation for a short period of time, 5 minutes). It is theorized that removal from a reinforcing environment decreases the probability of behavior which preceded it (Firster & Skinner, 1957). Time out is seen as a humane and nonaversive form of punishment with limited deleterious side effects. "Acting out" was operationally defined as "a set of disruptive behaviors which led to students receiving time outs, during their first 25 days in the program" (p. 18). Results indicated that Bender scores, based on six of Koppitz's emotional determinants in conjunction with the six previously elaborated on, correlated .462 for the combination of both sexes, .445 for females only, and .425 for males only. These Pearson Product Moment correlations were all significant at the .001 level. However, it is interesting to note that none of the Koppitz determinants, when used alone, reached any level of significance.

While this study is very provocative and well designed in a methodological sense, it remains correlational data. Such data is popularly understood to show a relationship or association between variables but in no sense can such a relationship be said to be indicative of
causation. The perpetual problem of confounding by a third variable as well as the notion that a correlational study cannot really specify which variable influences the other remains (Spence, Underwood, Duncan, & Cotton, 1968). Experimental research which involves active manipulation of a treatment variable is the only type of record which is able to establish evidence concerning the truthfulness of cause-effect statements. Such experimental manipulation of treatment variables has only rarely been performed with relation to influencing so called "personality tests" or "projective techniques" and none of the literature shows any direct evidence of this ever being attempted with regard to the Bender Gestalt test. Yet, the Bender may be one of the more sensitive indices of "personality" or response repertoire, which relates specifically to "acting out" behavior. The following study has investigated the degree to which behavior management procedures affect not only the rate of "acting out" behavior but also the extent to which these same procedures ultimately affected emotional indicators on the Bender Gestalt. Results from this study provide information concerning the viability of behavior analysis procedures as they affect what has been referred to by the more traditionally oriented psychologists as "personality." Further, the extent to which such "personality" indicators can be influenced by the manipulation of
environmental variables may remove some of the mysticism associated with the concept of "personality."

Method

Subjects

All subjects in experimental and control groups were classified as disruptive and tending toward "acting out" based on three criteria. One was the classroom teacher's observation of the child performing at least five "acting out" episodes per day. This determination will be assessed as a function of that teacher's need to intervene and curtail the child's "acting out" performance on each occasion. Such interventions may be physical or vocal. These events will be counted for 1 week and must average at least five events per day. The second criterion is the scoring of at least two Emotional Indicators on the Bender Gestalt Test. Previous to group assignment, the third criterion for placement in the control or experimental group will be assessed via the Devereux Elementary School Behavior Rating Scale (1967). The Devereux is a teacher rating procedure developed by the Devereux Foundation which uses a Likert-type scale to describe student functioning in the classroom. The Devereux measures a multitude of classroom behavior factors (11), one of which, classroom disturbance, is very reminiscent of the
behaviors typically described as "acting out." The specific questions for this factor which the teacher rates are as follows:

11. Has to be reprimanded or controlled by the teacher because of his behavior in class?

12. Pokes, torments, or teases classmates?

13. Annoys or interferes with the work of his peers in class?

30. Quickly drawn into the talking or noisemaking of others (i.e., stops work to listen or join in)? (1967, p. 3)

Questions 11, 12, and 13 are rated on a scale from 1 (never) to 5 (very frequently). Question 30 is rated from 1 (not at all) to 7 (extremely). A total of 20 points places the child beyond 2 standard deviations on this factor and as such (together with classroom observation data by teacher and psychologist, and the Bender) makes him eligible for placement in either the experimental or control group.

The test-retest correlations for the classroom disruption factor is .91, which is the highest of all factors on the test; the median reliability coefficient is .87 for all factors combined. All such statistical and practical information make the Devereux an excellent screening device for this study.
Subjects originally consisted of 24 elementary school children from a small urban school district who have shown the above diagnostic indicators of "emotional disturbance" and more specifically "acting out" behavior. Students who were randomly assigned to the experimental group must be reviewed by an administrative school committee known as the Admission, Review, Dismissal (ARD) committee. The committee determined the appropriateness of all placements into the forthcoming alternative classroom in which experimentalists were treated. Controls were exposed to the standard special education curriculum reminiscent of this particular district. Experimentalists and controls were divided into two groups of 10 students each.

Procedure

Bender tests were administered within the first 2 weeks of baseline for the experimental and control groups. Procedures for administration followed directives specified by Koppitz (1963, p. 15), but students were given a second sheet only in the event that it was specifically asked for. This follows the same procedural pattern as Caudle (1981) (see Appendix A). Scoring of these protocols was not performed until posttesting was completed and a second rater scored all protocols for reliability purposes. The subjects' names and group membership were identified by number only. Scorers were not aware of
the affiliation or group membership of the tests at the time of scoring. Percent agreements were determined by the formula: number of agreements divided by number agreements plus disagreements, multiplied by 100 (Agreement/Agreements + Disagreements) X 100.

As in the Caudle study, 12 emotional indicators were used as indexes of "acting out" behavior when scoring the Bender. These include six of Koppitz's emotional indicators and six others, previously mentioned, which are identified in the literature as indicative of "acting out." These in total include dashes substituted for circles; progressive increase in size for Figures 1, 2, and 3; large size; overwork or reinforced lines; second attempt; expansion; progressive increase in size from Figure A to Figure 8; sharp angles on Figure 6; collision; dashes substituted for dots; circles substituted for dots; and uneven figure size (Caudle, 1981). Any one of these above indices were counted as points toward a total score of emotional disturbance on the Bender. Maximum possible points for any one individual was 12 points. For further description of the diagnostic indicators, see Appendix B.

In order to show that an individual's repertoire, with regard to acting out as measured by the Bender, had been significantly affected by a treatment, a fairly rigorous and logical design was constructed. It is unlikely that,
even with the best of all possible experimental procedures, all competing explanation of change would be accounted for. With applied human research, one can rarely achieve the experimental rigor reminiscent of the laboratory. One can only hope to compensate for erratic environmental artifacts by using statistical controls.

One of the more suitable means of statistical control, where appropriate, is the analysis of covariance. This technique statistically adjusts posttest scores on the basis of one or more related variables, usually pretests. This is performed in an attempt to control existing group differences on those variables which influence the outcome previous to treatment. With regard to the dependent variable in this study, all subjects (experimental and control) have had their pretest scores on the Bender used as the covariate. To the extent that the number of emotional indicators on the Bender posttest can be predicted from performance on the pretest, such behavior could not be attributed to the treatment. The analysis of covariance permitted determination of the proportion of variances of the criterion (number of emotional indicators) existing previous to application of the treatment. This proportion was then removed from the final statistical analysis. Thus, statistical control was established in two forms. Error variance was substantially reduced and prevariation
in the criterion variable was controlled. In addition, in order to assess the effect of treatment, independent of comparison with the control group, a two-sample pre- and postintervention \( t \) test was performed on the experimental groups only. Like the analysis of covariance, the \( t \) test is robust with regard to violations of the assumptions relating to heterogeneity of variance when sample sizes are equal. Further, the \( t \) test gains control over potential threats to internal validity regarding selection bias and mortality. However, the \( t \) test does not compensate for threats to internal validity relating to history, testing, maturation, instrumentation, selection-maturation interaction, and regression as does the analysis of covariance. If the \( t \) test had proved statistically significant but not the analysis of covariance, the previously mentioned arguments for threats to internal validity would have had to be addressed and rebutted.

With this design in mind, subjects were randomly assigned to experimental and control groups on the basis of school location. Those subjects showing significant diagnostic indicators form two elementary schools within a small urban school district were placed in the experimental
group. Others showing the same indicators but from two other elementary schools in the district were assigned as controls. Controls received the standard intervention strategies reminiscent of the special education department of the school district. The only difference between the experimental and control group was the degree of formalized contingency management structure imposed on the experimental groups. Subsequent to pre-testing, the experimental group underwent 4 weeks of baseline observation. During this time the usual techniques regarding behavior management of obstreperous students were employed. One teacher, two teacher aids, and the psychologist endeavored to maintain order in the classroom and to teach according to the district's standard special education guidelines. During baseline (20 days), the frequency of each student's "acting out" behavior was recorded by the teacher. "Acting out" was operationally defined as each occasion during which a given child performed a behavior which was disruptive in nature. These inappropriate behaviors fell into the general categories of aggressive, noncompliant, and verbally abusive. In order for such behaviors to quality as data
for baseline, they had to be deemed serious enough by the teacher to warrant some corrective action.

Basically, data were obtained according to the same criterion specified in the Caudle (1981) study. In that study, "acting out" behavior was identified and defined as those events during which students' behavior becomes so disruptive as to warrant placement in time out. For the purpose of this study, the same definition applied, except that students were exposed to other corrective modalities beyond the use of time out. Data consisted of the number of seriously inappropriate behaviors that occurred during the total 6-hour school day which required corrective intervention. Intervention was defined here as the teacher, teacher's aid, or psychologist taking one of the following steps:

1. Warning in low tone (verbal prompt in low tone that the next incident of disruption will result in time out), or

2. Placement in time out.

Either of the above was counted as a discrete event reminiscent of "acting out" behavior and counted as part of the frequency of inappropriate behavior performed per day. Such data were taken on each student (O'Leary & O'Leary, 1972). Each of these instances is readily recognizable to the teacher or psychologist and as such
makes an easy criterion for continuous event counting (Martin, Garry, & Pear, 1978). Other potential possibilities for data keeping included time sampling, partial interval recording, and whole interval recording. But these systems of data keeping have recently been described as seriously flawed in research described by Springer (1981). With regard to discontinuous time sampling, it was found that for time intervals over 40 seconds, the data tended to overestimate the actual percentage of behavior occurring during a given period of time. Likewise, partial interval recording tended to overestimate responding while whole interval recording tended to underestimate true ratios of behavior being emitted during a given session (Powell, Martindale, & Kulp, 1975).

Inter Observer Reliabilities (IOR) was determined by sampling teacher data observations concurrent with a second observer during selected intervals. The smaller of the two total intervention events taken during a 30-minute interval was divided by the larger number obtained by the second observer during that same period. The quotient was then multiplied by 100 to give the IOR percentage reliability. Though this strategy left open the possibility of occasional uncorrelated observations, large agreement percentages tend to lend confidence to the actual number of necessary intervention events during a given session.
(Martin & Pear, 1978). Further, this tactic is consistent with recommendations made by Springer (1981) in which she states:

Since continuous measurement does not necessarily require any more observation time than discontinuous measurement, it offers the possibility of more information, in terms of accurate descriptions of response dimensions, at no increase in cost. (p. 26)

The "acting out" data which were counted for each individual subject (experimental and control) were plotted across the 20 days of baseline and treatment and then subjected to time-series analysis. This was also performed on total group data for experimentals and controls. The purpose of analyzing the frequencies of "acting out" behaviors prior to and subsequent to treatment is to further establish the effect of the treatment modality and to determine to what extent outcomes, in terms of the final number of Bender Emotional Indicators, correspond to the frequency of actual "acting out" behavior.

Statistical significance in the changes within each individual's curves, across phases of the experiment, was computed via time-series analysis. Other possible single-subject procedures such as the single-subject analysis of variance models introduced by Shine and Bower
(1971) and Gentile, Roden, and Kelin (1972) were rejected, as they assume the single-subject to be a response generator whose responses are statistically independent (Glass, 1975, p. 78). But as Glass (1975) points out, operant research generates data in which nonindependence of observations is a highly probable phenomenon. Therefore, time-series analysis was selected as the appropriate data analysis because it is most consistent with operant assumptions.

Time-series involves use of "successive observations throughout a programmed intervention and assesses the characteristics of the change process" (Gottman, McFall, & Barnett, 1969, p. 299). Gottman et al. propose that this design serves several simultaneous advantageous functions. First, it provides a descriptive function because a continuous record of the experimental variables are utilized over the entire time period. Secondly, this design serves as a useful heuristic device because the time-series data provide feedback for generating new hypotheses or as a source for post hoc hypotheses. Thirdly, it can serve as a quasi-experimental design for planned intervention in a total program without the necessity of a control group. These authors propose that the use of time-series is better at ruling out rival hypotheses than the one-group pretest/posttest design, and more thoroughly enables
examination and hypothesis testing about the process of change.

It is only after the correct model for each individual subject is determined via the correlogram that determination of intervention effects can be computed (Glass, 1975, p. 74). Establishing the correct model for each subject's data analysis was accomplished by three procedures. First, the autoregressive process \((p)\) regresses upon itself one time point so that any time points are predictable from the observation of previous data. This technique is used by Glass (1975) because of his determination that no particular time point is independent of preceding time points. Thus, such an interdependence can be used to forecast future values. This autoregression is a correlation procedure used to determine the degree to which time is a relevant variable or merely a random fluctuation of data points. Second, the order of differencing \((d)\) is a method of determining what differences in the previous time period recur. Third, the order of the moving averages \((q)\) addresses itself to data trends which are designated as random shocks which enter the system. This whole procedure ultimately measures a trend over a given time interval and allows for smoothing of the time line.
Glass (1975) points out that about half of the time-series found in practice are adequately described by stationary models. That is, the series tends to maintain an equilibrium around a constant mean level. When this is found to be the case, via examination of the correlogram, $p$ is given an order of 1 in the $p$, $d$, $q$ series. When $p$ is 2, the autoregression tends to depart from linearity and assumes some other form such as a quadratic or cubic trend. When $p$ is 0, a stationary trend does not exist, and $d$ and $q$ must be examined as further indicators of possible trends.

With regard to $d$, a series is stationary (assumes a limited degree of fluctuation around a particular level) when $d$ is 0. A $d$ of 1 indicates a stationary trend at a particular level for a given time, then drifting to a new level where it again assumes a stationary trend. When $d$ is 2, the change in level of the stationary trend will drift from location to location on the correlogram.

When $q$ is 0, the process is purely autoregressive and no moving average exists. When $q$ is 1, there is a first order moving average; and when $q$ is 2, there exists a second order moving average (Glass, 1975).

After model specification was determined, a $t$ test was used to decide if there were any significant differences in the regression estimates of the trend.
lines (baseline and experimental) for each individual subject (Glass, 1975, p. 119).

Since obtainment of significant differences in time-series analysis is most critical at the point just after intervention, a delayed treatment effect, reminiscent of most learning curves, reduces power. However, Glass (1982) concedes the need to extend the point of the time-series test to some reasonable period of time after treatment has been in progress. Since teaching disruptive children to behave appropriately in the classroom is definitely a gradual process and since reduction of "acting out" behaviors reflects this process of learning, the point of introducing the time-series test was extended 2 months posttreatment such that only the last 20 days of treatment were compared to the first 20 days of baseline. The number of "acting out" behaviors per day was then compared in terms of frequency per day during the first and last phases of experiments. Glass notes that this strategy of dividing the time-series into two separate segments introduces error into the assumption of consistent dependence between adjacent samples, but he indicates that time-series analysis is probably robust with regard to this assumption. Further, for reason of a more practical nature, it was probably not possible to have teachers of control subjects, outside the alternative
classroom, take data on each child throughout the course of the entire semester.

Treatment

The notion that behavioral strategies have a beneficial effect on the overall personality and emotional well being of the emotionally disturbed child has, at least, received tacit if not empirical support throughout many school districts. And it is popularly understood that many, if not most, school psychologists spend a great deal of time involved in personality testing (O'Leary & O'Leary, 1972). Subsequent to personality assessment it is frequently the case that at least part of the emotionally disturbed individual's care and treatment may consist of behaviorally oriented intervention. Operant strategies have gained increasingly in popularity throughout many school districts in past years. The degree to which this treatment has a pervasive effect such that much of the child's repertoire (personality) is altered is a function of the type and perhaps even the duration of that treatment.

The treatment which was applied in this alternative classroom setting consisted of variations on classroom management (O'Leary & O'Leary, 1972), token economy (Allyon & Azrin, 1965), contingency management (Fenley, 1979), and direct instruction (Engleman,
Carnine, & Johnson, 1978) themes. The paradigm is basically operant in scope and emphasis. The strategies involved the teaching of on task behaviors which were contingent on a menu of reinforcers.

One teacher, two teacher aids, and a psychologist administered this program to a class of children who had a well documented history of "acting out" behaviors. As stated earlier, these were children who represented the most severe discipline problems in four elementary schools. The attempt was to shape various components of on task behavior by reinforcing approximations of this behavior. On task was operationally defined as having the child perform requisite behavior related to completion of academic tasks and appropriate peer interactions. This usually required that the child be seated at his desk with both feet on the floor, performing academically oriented behavior or, while in free time, playing constructively with other children. When such behavior occurred according to the above description, the student received tokens (Peabody chips) which were accumulated and then spent on various items from a menu of reinforcement. Such a menu consisted of an opportunity to spend more free time in the Reinforcing Events (RE) area playing with various games and recreational facilities or of an opportunity to spend chips at the end of each hour on a very few, limited number of primary reinforcers (unsweetened
crackers and fruit juice). Administration of such tokens was, of necessity, on a variable interval schedule and subject to the teacher's, psychologist's, and aids' discretion. At various times the reinforcement density was dense and at others, depending on the rate of on task behavior, such reinforcement density was very lean.

A somewhat revolutionary notion expressed by Sprick (1981) is that as a child becomes more disruptive, it becomes more critical to reinforce existing on task behavior in conjunction with corrective strategies for off task, disruptive behavior. Only as the child shows greater indications of stimulus control can the teacher or psychologist extend the requirements of the reinforcement schedule.

Dispensation of tokens occurred on approximate 3-minute intervals during the first weeks of the program and gradually faded to 10-minute intervals as on task behavior increased. When tokens were exchanged for free time, the rate of exchange was always one token for 1 minute of free time. However, usually it was necessary for a student to accumulate four or five tokens before being allowed access to free time as a reinforcer. Primary reinforcers (crackers, etc.) were exchanged one for one. Social praise was usually given in conjunction with tokens or tangible reinforcers.
On isolated occasions, when individuals were particularly obstreperous, the use of time out from reinforcement was applied. Time out from reinforcement was defined here as the removal of an offending disruptive student from the classroom and his subsequent placement in an isolation for a period of 3 to 5 minutes or until such time as the individual was calm and not involved in any overt hostility (Sprick, 1981). The attempt was to use this strategy only as a last resort when the disruptive individual had not responded to a low tone verbal warning (O'Leary & O'Leary, 1972).

The obvious question at this point became how do these procedures effect the outcome of Bender Gestalt protocols. By way of a circuititious explanation, it should first be mentioned that Koppitz (1963) agrees with Bender (1938, p. 157) regarding the notion that neurotic disturbances do not result in aberrations of perception or visual motor function. Nor is it the case that perceptual problems are necessarily related to emotional unrest. As stated earlier, indices on the Bender can differentiate between perceptual problems and emoitional indicators (Clawson, 1959; Eber, 1958; Pascal & Stutlell, 1951). As Simensen (1974) points out, children with severe emotional problems are not different from normal children with regard to perception and coordination but they
display maladjustments in the form of integration of the two (Koppitz, 1963). These children are so distorted in orientation to their environment, in such a state of continual confusion, that they cannot attend properly to the simple task of copying a design. These children show malfunction, according to Koppitz, in visual-motor perception as a function of their emotional disturbance. A person who demonstrates emotional disturbance is sometimes described as lacking self-confidence and self-control. These are at least several of the many features which fall into the rubric (Koppitz, 1963). It is anticipated that to the extent that positive interactions can be increased between classroom staff (teachers, psychologists, and teacher aids) and the children, as well as among the children, these students will tend to display behaviors which our culture tacks as "self-confidence." Sprick (1981) notes that

A poor self-image manifests itself in many different behavior problems. Acting out typically indicates a poor self-image. This is behavior custom-designed to get the teacher's attention. A child who sees himself as a good student will be a good student. A child who sees himself as stupid will fail in academic endeavors. (p. E9)
It might be said that the child's inclination to say negative things to himself about himself determines, to a large extent, his inability to succeed. The child's negative self-repertoire is manifest in many forms. Since we as observers of overt behavior do not have access to the covert verbalizations of the child we must determine some correlary to the child's "inner disturbance." Since "acting out" is often seen to be representative of such phenomena, the degree to which we can decrease the child's propensity to such distractive behaviors via positive interactions with teachers and peers may, to some extent, effect that behavior, metaphorically referred to as "poor self-concept." Further, it was hypothesized that any effect on the child's positive interactions which are reflected in terms of increased on task behaviors and increased socialization may also be reflected via a test which measures various indicators of emotional disturbance. Obviously, emotional disturbance, even when it is related to what has been called a "poor self-concept," is not solely under the influence of the teacher. There are many people who shape the child's "personality" or social repertoire. But the classroom teacher is the person who probably gives the greatest amount of feedback to the child during the course of an average day.
Schoolwork, during childhood, is by far the individual's most important endeavor and perhaps his greatest challenge. The personal interactions the child has with his classroom teacher may be one of the most critical influences in the child's development of what we call self-concept. (Sprick, 1981, p. El)

As specified earlier, the Bender Gestalt test has been found to correlate well with children's propensity toward "acting out" and is also often seen as an overall index of emotional disturbance (Koppitz, 1963). If all of these premises can be accepted, it is a small inductive leap to anticipate that increasing on task behaviors and positive teacher-student and peer interactions will eventually be measured in improved indicators of emotionality on the Bender.

Unfortunately, it is very difficult to be about the business of teaching a class of 10 emotionally disturbed children and be constantly measuring all behaviors reminiscent of positive interactions and on task behavior. A much more salient series of behaviors which the teacher could readily take note of was the number of negative interactions and conspicuous disruptive off task behaviors. This is simply the reciprocal of the former. These behaviors are not only salient but have been shown to
correlate with the emotional indices on the Bender. As previously stated, it was hypothesized that a decrease in "acting out" behaviors, which occur as a function of differentially reinforcing increased positive interactions among students and between staff and students, will, because of common contingencies, eventually be reflected in some significant decrease in the number of indicators of emotional disturbance on the Bender Gestalt.

It is hoped that many of the contingencies which inadvertently led to the child's emotional problem were reversed. Many of the children who were, previous to treatment, acknowledged only when they caused disruption were, as of the beginning of treatment, acknowledged primarily when their behavior was appropriate. Simple off task behaviors were ignored. Disruption was not attended to; it was extinguished in time out or simply resulted in a low tone warning of forthcoming time out. Though all of these tactics were counted for data purposes, they did not constitute negative attention reminiscent of most classrooms. Positive, on task behaviors and interaction behaviors were given special attention via verbal and token reinforcement on a variable interval schedule (VI 3' to VI 10'). Classroom curriculum was adjusted to the student's level of function and one to one tutoring was given at various intervals.
all through the school day. Direct instruction materials (Engleman et al., 1978) were used to promote group coordinated response and remediation in reading deficiencies. Tokens were exchanged for access to a free time area where small groups of students (three to five at one time) interacted with various toys and games. Staff further reinforced appropriate peer interaction during this free play time. These may be important components that relate indirectly to increased precision when copying the Bender design. Many of the children in both the experimental and control groups were rather poor at following any directions. They often failed to attend to simple directions and failed to comply even when attending behavior seemed apparent. The process of verbally and tangibly reinforcing approximations of on task behavior, both in direct instruction and when working individually, may have influenced the way in which directives are followed on subsequent Bender tests.

The general effort was to increase positive praise statements contingent on improved class behavior and individual performance. Noncontingent praise was not given. Sprick (1981) indicates that most classroom teachers operate at about a 3 to 1 negative to positive comment ratio while teaching. He recommends reversing this ratio of 3 to 1 in favor of positive comments. Every
effort was made to fulfill this recommendation. Sprick indicates that this strategy alone has proven to be extremely effective in influencing the "self-concept" of disruptive students.

Summary

To summarize, the basic intervention strategy to be used with these disruptive children was a global operant classroom contingency management procedure. Teachers, aids, and the psychologist gave tokens and praise for increased approximations of on task behavior and also reinforced appropriate student social interactions. Inappropriate behaviors resulted in the student being warned that time out was forthcoming if he persisted and actual placement in time out if the offending student continued to misbehave. The number of occasions during which each student was presented with a warning or actual placement in time out was one source of data to be analyzed to determine the effectiveness of the treatment. It was hypothesized that there would be a reduction in the frequency of such behaviors over the time of the total semester. Further, the significance of such a reduction was computed via time-series analysis. Baseline consisted of the first 20 days of class and the treatment was measured during the last 20 days. This procedure is consistent with a recommendation by Glass (1982) and also allows
acquisition of similar data for control subjects in other classes throughout the district. It is important to note the criterion for measurement of inappropriate behavior with controls is basically the same as the experimental.

A more subtle, but perhaps controversial, determination of effect on the global repertoire or "personality" of these students was examined via a pre- and posttesting for experimental and controls by use of the Bender Gestalt Test. The number of Emotional Indicators, pre and post, was the criterion for change. The analysis of covariance was used to test this outcome. Specifically the null hypothesis ($H_0$) indicated that means of both groups are equal with regard to the number of Emotional Indicators on the Bender Gestalt when they are adjusted for pretest scores. The alternative hypothesis ($H_1$) stated that the means of the experimental group ($U_1$) would be less than control group ($U_2$) when they were adjusted for pretest scores.

In addition, a pretest-posttest design was performed on the experimental group only. A $t$ test determined the effect of treatment independent of comparison with the control group. The null hypothesis ($H_0$) stated there would be no difference between the means of pretest and posttest on the Bender Gestalt.

**Results**

The previously described procedure was accomplished as prescribed. Subjects in elementary special education, both
experimental and controls, were exposed to contingency management strategies for the duration of the spring semester. Data acquisition was performed in an unobtrusive and efficient manner.

Inter observer reliabilities on the number of acting out episodes varied across teachers from a low of .83 to a high of 1.00. On most occasions reliability scores were a perfect one to one matching between teacher and psychologist or between teacher and teacher aid. This high reliability occurred simply because verbal correction or placement in time out are fairly salient and conspicuous features even to the most casual of observers. It is important to note that the criterion for an "acting out" episode was not simply the occurrence of an inappropriate behavior on the part of the student. In order for the behavior to be counted as an occurrence of "acting out" it must have been judged serious enough for the teacher, psychologist, or teacher aid to have intervened either by verbal warning or by placement in time out. Though such episodes are indeed conspicuous events and subject to easy verification, the possibility for observer drift remained. Observer drift occurs when the criterion for correction (i.e. the severity of the inappropriate behavior) changes across time. In the case of improving students the observers' criteria could become more rigorous. In the case of an increasingly belligerent
class the criteria could become more lax. In order to
gauge the possibility for such a threat to reliability the
psychologist made note of the initial and terminating
episodes that resulted in verbal warnings and/or time out
for experimental and control students. Criteria in both
groups remained relatively constant across time. Typical
behaviors which were corrected by teachers included talking
out, out of seat behavior, hitting and other types of physi-
cal aggression and noncompliance. These behaviors, when
they occurred, appeared to be constant in intensity whether
at the beginning or end of the term. Such behaviors are,
again, fairly salient; when students were out of their
seats or talking out, such events were unlikely to go un-
noticed or uncorrected.

Reliabilities for noting the number of Emotional
Indicators on the Bender Gestalt protocols were also high.
A second rater judged the protocols without benefit of
knowing student group or identity. Criteria for judging
an Emotional Indicator on the Bender are also straight-
forward and can be seen in Appendix B. The reliability
coefficient for the Bender was calculated at .87.

It is important to note that 2 of the experimental
subjects moved from the district during the course of the
experimental semester. This reduced the number of experi-
mental subjects from 12 to 10. In order to maintain a
balance between groups, 2 of the control subjects were also
randomly dropped from the study. Intervention with these eliminated controls was, however, maintained independent of data acquisition.

Hypotheses

The null hypothesis for the analysis of covariance (number 1) stated no posttest differences in the mean number of Emotional Indicators on the Bender Gestalt when the experimental and control groups were adjusted for pretest scores. The F ratio as calculated by the Statistical Package for the Social Sciences was 4.498. The critical F ratio at the .95 significance level with 1 and 17 degrees of freedom is 4.45. Therefore the null hypothesis of no difference is rejected at the alpha level of .05 (p < .05). See table 1. The research hypothesis for the analysis of covariance (number 2) stated that there would be a difference in these group means when adjusted for pretest differences. The research hypothesis for analysis of covariance is therefore accepted.

It will be quickly noticed that the critical F ratio of 4.45 is very close to the obtained F ratio of 4.498. This should not be interpreted as a minimal effect with regard to the treatment. The posttest mean number of Emotional Indicators dropped for both groups. The experimental group dropped from 3.2 to 1.5, a difference of 1.7, while the control group reduced their average number of Emotional Indicators from 2.6 to 1.9, a difference of .7. The reason
Table 1

Analysis of Covariance for Experimental and Control Groups on Pretest and Posttest Bender Gestalt Emotional Indicators of "Acting Out" Behavior

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>$SS_x$</th>
<th>SP</th>
<th>$SS_y$</th>
<th>df'</th>
<th>$SS'_y$</th>
<th>$MS'_y$</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>1</td>
<td>1.8</td>
<td>-1.2</td>
<td>.8</td>
<td>1</td>
<td>2.195</td>
<td>2.195</td>
<td>4.49*</td>
</tr>
<tr>
<td>Within</td>
<td>10</td>
<td>22</td>
<td>10.6</td>
<td>13.4</td>
<td>17</td>
<td>8.293</td>
<td>.488</td>
<td>xxx</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>23.8</td>
<td>9.4</td>
<td>14.2</td>
<td>18</td>
<td>10.49</td>
<td>xxx</td>
<td>xxx</td>
</tr>
</tbody>
</table>

*P < .05
for a mutual drop in experimental and controls is most probably accounted for by mutual treatment rather than regression or maturation. Control subjects as well as experimental subjects were exposed to the influence of behavioral management strategies. Controls were not permitted to simply forego intervention. Ethical considerations required complete psychological intervention to the extent possible within the confines of the existing district framework. The only difference between the groups was in the extent to which teachers were able (and perhaps willing) to follow the directives of the psychologist. The teacher and aids for the experimental group were picked specifically for their ability to carry out behavior-management procedures and for their eagerness to work in conjunction with the psychologist. Another possible advantage for the experimental group was the increased student-to-teacher ratio. This ratio of teachers to students fluctuated between 4 to 12 and 3 to 8 for the experimental group, while in most cases controls averaged only 1 and sometimes 2 teachers to about 10 or 12 students. The \( t \) test performed for dependent observations on the experimental group alone illustrated the degree of experimental effect in a more pronounced and conspicuous fashion.

The null hypothesis (number 3) for the \( t \) test on correlated observations stated no differences in the
forthcoming pretest and posttest means of emotional indicators on the Bender Gestalt. The $t$ test required a ratio of 4.59 at the .001 level of significance ($4.59 = 4.59$). The obtained $t$ ratio was calculated at 5.605 which is highly significant at the .001 ($p < .001$); the null hypothesis was therefore rejected. However, level of significance is not necessarily indicative of degree of effect. To determine the extent of mean difference from zero a confidence interval is required. The .001 confidence interval is $1.7 \pm 1.39$ or a lower limit of .31 and an upper limit of 3.09. There is thus 99.9% assurance that the retest mean exceeds the population mean of the initial test by some amount between .31 and 3.09. The research hypothesis (number 4) stating a forthcoming difference between the pretest and posttest is therefore accepted. This $t$ ratio is a much better index of degree of treatment effect than the before-mentioned analysis of covariance. Though the $t$ test is subject to more threats to internal validity it has the benefit of a comparison made with the experimental subjects alone, independent of contaminating treatments performed on control subjects.

The preceding two analyses (anova and $t$ test) show significant decreases in the number of Emotional Indicators on the Bender Gestalt of emotionally disturbed children between the beginning and end of the spring 1982 semester. Changes in the number of such indicators are believed to be
a function of a behavior modification treatment performed during this time. In order to further substantiate the influence of this behavioral treatment, the frequency of "acting out" behaviors will be analyzed to detect changes in frequency over the course of this semester. It will be recalled that the Bender Gestalt Emotional Indicators selected for analysis were those reminiscent of such "acting out" behavior.

Research hypothesis (number 1) indicated a decrease in the frequency of "acting out" behaviors for the experimental group as a whole. Once again, "acting out" was operationally defined as behavior performed by the student/subject which was so inappropriate as to require corrective intervention on the part of the staff. Corrective intervention consisted of a warning of forthcoming time out or actual placement in time out. The frequency of such "acting out" behaviors for each individual were totaled across groups such that the total group frequency per day could be analyzed for group changes in frequency. Although there was disparity between the entering dates of some experimental subjects, the daily group totals serve as a good approximation of group behavior during the first 20 days of baseline and the last 20 days of treatment. Time-series analysis correlations performed on the experimental group totals for this time period indicated a moving average model (0, 0, 1). The
time series experiment produced an error variance of 68.231. The change in level was -43.01 with a \( t \) test value of -16.78. This was highly significant at the .001 (\( p < .001 \)), and the null hypothesis of no differences in frequency was rejected while the research hypothesis of a change in frequency was accepted for the experimental group as a whole.

Figure 1 shows a relatively stationary group trend of high baseline rates of "acting out" behaviors that were reduced to almost half by the last 20 days of the treatment. See table 2 for a complete summary of the time-series intervention analysis of both group totals and all individual subjects within each group.

The research hypothesis for the control group (number 3) indicated no expected differences in group frequencies between the first and last 20 days. However, as previously mentioned, ethical considerations required treatment of controls although the circumstances for such treatment were often not as favorable as those for the experimental group. Contrary to expectations, the control group total frequencies across subjects indicated a significant decrease in the frequency of "acting out" behaviors. The correlogram model for this group was a moving average (0, 0, 1) and the error variance for the time-series experiment was 59.968. Change in level was specified at -22.41 with a \( t \) test value of -8.67 which is highly significant with a \( p \) of less than .001.
Fig. 1—The total number of daily "acting out" episodes for the Experimental Group during the first 20 baseline days and the last 20 days of treatment.
Table 2
Summary Table for Time-Series Intervention Analysis

<table>
<thead>
<tr>
<th>Identification</th>
<th>df</th>
<th>Time-Series Model</th>
<th>Error Variance</th>
<th>Change in Level</th>
<th>Significance of Intervention</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Group</td>
<td>38</td>
<td>0, 0, 1</td>
<td>68.231</td>
<td>-43.01</td>
<td>-16.78</td>
<td>.001</td>
</tr>
<tr>
<td>Control Group</td>
<td>38</td>
<td>0, 0, 1</td>
<td>59.966</td>
<td>-22.41</td>
<td>-8.67</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 1</td>
<td>38</td>
<td>0, 0, 1</td>
<td>8.882</td>
<td>-5.69</td>
<td>-7.09</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 2</td>
<td>38</td>
<td>0, 0, 1</td>
<td>12.474</td>
<td>-5.31</td>
<td>-5.36</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 3</td>
<td>38</td>
<td>0, 0, 1</td>
<td>8.368</td>
<td>-5.15</td>
<td>-4.99</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 4</td>
<td>38</td>
<td>0, 0, 1</td>
<td>1.299</td>
<td>-2.16</td>
<td>-5.68</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 5</td>
<td>38</td>
<td>0, 0, 1</td>
<td>5.742</td>
<td>-4.70</td>
<td>-6.20</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 6</td>
<td>38</td>
<td>0, 0, 1</td>
<td>3.308</td>
<td>-2.66</td>
<td>-3.64</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 7</td>
<td>38</td>
<td>0, 0, 1</td>
<td>10.498</td>
<td>-7.12</td>
<td>-7.99</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 8</td>
<td>38</td>
<td>0, 0, 1</td>
<td>1.509</td>
<td>-2.30</td>
<td>-5.33</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 9</td>
<td>38</td>
<td>0, 0, 1</td>
<td>3.147</td>
<td>-2.56</td>
<td>-3.86</td>
<td>.001</td>
</tr>
<tr>
<td>Experimental 10</td>
<td>38</td>
<td>0, 0, 1</td>
<td>4.644</td>
<td>-5.54</td>
<td>-6.88</td>
<td>.001</td>
</tr>
<tr>
<td>Control 11</td>
<td>38</td>
<td>0, 0, 1</td>
<td>6.016</td>
<td>-.19</td>
<td>-.26</td>
<td>--</td>
</tr>
<tr>
<td>Control 12</td>
<td>38</td>
<td>0, 0, 1</td>
<td>5.853</td>
<td>+.41</td>
<td>+.46</td>
<td>--</td>
</tr>
<tr>
<td>Control 13</td>
<td>38</td>
<td>0, 0, 1</td>
<td>3.217</td>
<td>+.23</td>
<td>+.32</td>
<td>--</td>
</tr>
<tr>
<td>Control 14</td>
<td>38</td>
<td>0, 0, 1</td>
<td>7.298</td>
<td>-3.16</td>
<td>-3.39</td>
<td>.01</td>
</tr>
<tr>
<td>Control 15</td>
<td>38</td>
<td>0, 0, 1</td>
<td>3.768</td>
<td>-3.52</td>
<td>-7.57</td>
<td>.001</td>
</tr>
<tr>
<td>Control 16</td>
<td>38</td>
<td>0, 0, 1</td>
<td>13.169</td>
<td>-7.08</td>
<td>-5.07</td>
<td>.001</td>
</tr>
<tr>
<td>Control 17</td>
<td>38</td>
<td>0, 0, 1</td>
<td>9.526</td>
<td>-3.42</td>
<td>-3.86</td>
<td>.001</td>
</tr>
<tr>
<td>Control 18</td>
<td>38</td>
<td>0, 0, 1</td>
<td>3.939</td>
<td>+.77</td>
<td>+1.33</td>
<td>--</td>
</tr>
<tr>
<td>Control 19</td>
<td>38</td>
<td>0, 0, 1</td>
<td>6.778</td>
<td>-5.46</td>
<td>-6.40</td>
<td>.001</td>
</tr>
<tr>
<td>Control 20</td>
<td>38</td>
<td>0, 0, 1</td>
<td>1.067</td>
<td>-.34</td>
<td>-.77</td>
<td>--</td>
</tr>
</tbody>
</table>
Figure 2 illustrates the substantial drop in group frequencies of "acting out" behavior for the control group totals. This corresponds somewhat to the drop in control group mean Emotional Indicators on the previously described Bender Gestalt analysis via the analysis of covariance. However, as with the Bender Gestalt results, control group changes are a function of great individual variation. Some of these students responded well to therapeutic procedures. Others, due to circumstances somewhat beyond the control of the researcher, did not show such favorable changes. It is only by analyzing the individual time-series data that this heterogeneous picture becomes clear.

The research hypotheses for all individual time-series analyses in the experimental sample anticipated significant decreases in "acting out" frequencies for each individual within that group. This research hypothesis was supported for all members of the experimental group. Correlogram data was identified as showing variation reminiscent of a moving average (0, 0, 1) for all subjects within that group. The following will elaborate on frequency changes in "acting out" behaviors in conjunction with the number of Bender Emotional Indicators on pretest and posttest.

Experimental subject 1 was by far the most seriously disturbed and obstreperous student in either of the groups. His pretest Bender Gestalt protocol showed 7 indicators of
Fig. 2—The total number of daily "acting out" episodes for the Control Group during the first 20 baseline days and the last 20 days of treatment.
emotional disturbance. His posttest protocol still showed a total of 3 such indicators. During baseline he showed the greatest number of inappropriate (often aggressive) behaviors requiring a warning or placement in time out. There was, as can be seen in figure 3, great variation in the frequency of such "acting out" from day to day. On several days he was seen to "act out" on as little as 3 occasions; yet on a given day during this period he managed to behave inappropriately a total of 16 times. The last 20 days of treatment still showed a considerable inclination toward inappropriate behavior but the frequency of these occasions had dropped considerably. The last days of school were particularly difficult for this child. He began to display many of the behaviors reminiscent of baseline. This was perhaps, in part, due to a series of family episodes which exacerbated classroom intervention. Error variance for this subject was 8.882 and level change was -5.69. This resulted in a $t$ test value of -7.09 which was highly significant with a $p$ of less than .001. Though statistical significance was evidenced by the above figures, the reader should make no assumptions regarding the overall change of this subject. His "acting out" rate at posttest as well as Bender protocol analysis was still indicative of severe disturbance.

Experimental subject 2 demonstrated inappropriate behaviors somewhat more frequently than his Bender Gestalt
Fig. 3--The total number of daily "acting out" episodes for Experimental subject number 1 during the first 20 baseline days and the last 20 days of treatment.
protocols might suggest. He scored only 2 Emotional Indicators on the pretest and one on the posttest. However, the frequency of his baseline "acting out" was high and erratic (see figure 4). During treatment he demonstrated considerable verbal and physical resistance and came under verbal control of the teacher only toward the end of the semester. This student had a long history of verbal and physical abuse and medical reports indicated considerable neurological tissue damage as a function of such previous abuse. By the end of the semester he began to show consistent signs of compliance but had occasional episodes of aggression toward peers and teacher aids. Time-series analysis indicated an error variance of 12.474 and a change in level of -5.69. His t-test value was -7.09 demonstrating a significant p of less than .001. This subject was seen by many at the initial stage of therapy as having one of the worst prognoses in the group. However his frequencies of "acting out" during the last 20 days of treatment and his Bender Emotional Indicator count of 1 seem to show that he made considerable improvement during this time period.

Experimental subject 3 was perhaps the most mischievous but was seen by the staff as the least disturbed student in that group. His pretest Bender protocol exhibited only 2 Emotional Indicators and his posttest dropped to only 1. He demonstrated a relatively high baseline rate of "acting out"
Fig. 4—The total number of daily "acting out" episodes for Experimental subject number 2 during the first 20 baseline days and the last 20 days of treatment.
throughout the course of the first 20 days but by the end of the semester his rate had dropped to almost one-third of his initial rate (see figure 5). He nevertheless persisted in frequent attempts at haranguing other students and generally playing the role of instigator right up till the end of the term. Error variance for the time-series was 8.368 with a change in level of -5.15. His t test value was -4.99 resulting in a significant p less than .001.

Experimental subject 4 was a comparatively quiet child who only showed erratic bursts of noncompliance during baseline. Such noncompliance was frequently of a passive nature and her most typical type of inappropriate response consisted of aimless out-of-seat behavior. Her pretest Bender showed 3 Emotional Indicators and this dropped to only 1 by the time of posttest. Error variance for her time-series was 1.299 with a level change of -2.16. This produced a t test value of -5.68 which was significant with a p of less than .001. During the last 20 days of the semester her behavior became so compliant and generally amiable that no "acting out" behavior was noted on a regular basis. During the last 2 weeks of school this subject was moved to a less restrictive classroom for learning disabled children so that more concentration might be placed on those in the experimental class who were still demonstrating "acting out" behaviors. The teacher in the learning disabled classroom
Fig. 5.—The total number of daily "acting out" episodes for Experimental subject number 3 during the first 20 baseline days and the last 20 days of treatment.
took up the job of documenting any episode of acting out, but as can be seen in figure 6 these were very rare instances. Perhaps this subject was one of the best examples of the potential benefits to be derived from such an experimental class.

Experimental subject 5 had 4 Emotional Indicators on his Bender pretest and showed only 2 on his posttest. The inappropriate behaviors he demonstrated were seldom aggressive or hostile; more often he was noncompliant and inclined to wander about the room off-task. He was also somewhat prone toward talking out without permission and frequently demonstrated irrelevant commentaries. His level of scholastic functioning was particularly low. He demonstrated an erratic rate of "acting out" behaviors during the baseline period, but this rate dropped markedly during the last 20 days of treatment (see figure 7). However, even during the last phase of the semester he maintained a somewhat constant daily "acting out" level, averaging a little over 2 episodes per day. Error variance for this subject was found to be 5.742. Change in level was -4.70. The t test value was -6.20 which was highly significant with a p of less than .001.

Experimental subject 6 demonstrated an almost average level of academic functioning but performed a wide variety of off-task behaviors in order to avoid academic compliance. Though not particularly aggressive, he often refused to work
Fig. 6—The total number of daily "acting out" episodes for Experimental subject number 4 during the first 20 baseline days and the last 20 days of treatment.
Fig. 7--The total number of daily "acting out" episodes for Experimental subject number 5 during the first 20 baseline days and the last 20 days of treatment.
or was randomly out of his seat during baseline. The rate of these behaviors, by experimental class standards, was not particularly high. He averaged about 3 "acting out" episodes per day during baseline but dropped to between 0 and 1 by the last 20 days of treatment. Toward the end of the semester his behavior had improved to the extent that it was deemed feasible and appropriate to allow him to partake of regular, "mainstream" class activities for 1 hour per day. This hour was monitored for "acting out" episodes also.

His pretest Bender protocol evidenced only 2 Emotional Indicators and dropped to 1 by posttest. Error variance on time-series was 3.308 with a level change of -2.66. The t test value was -3.64 which was significant at a p of less than .001. (See figure 8.)

Experimental subject 7 was perhaps another whose Bender protocol failed to demonstrate the degree of severity in her emotional disturbance. This subject had a long history of physical abuse and occasionally came to class with open sores. Her pretest Bender showed only 3 indicators of emotional disturbance, but the experimenter is inclined to believe that this is not an adequate representation of degree of severity at that time. By posttest her Emotional Indicators had dropped to 2, but this again is perhaps somewhat incongruous with her overt behavior. Baseline frequencies of acting out averaged around 10 per day but dropped
Fig. 8--The total number of daily "acting out" episodes for Experimental subject number 6 during the first 20 baseline days and the last 20 days of treatment.
to about 3 by the last 20 days of treatment. She nevertheless continued to show erratic episodes of aggression and noncompliance and made dramatic attempts to avoid time out when directed to that location. Though the psychologist believes that some preliminary change has taken place in her behavior, there remains much work to be done with this child. Error variances for the time-series was 10.498 and the level change was -7.12. The t test value was -7.94 which gave a significant p less than .001. (See figure 9.)

Experimental subject 8 was a case which demonstrated the reverse indicators of the former subject. Though he showed 3 Emotional Indicators on the Bender pretest and only 1 on the posttest, his classroom behavior never actually became difficult to control. Even on baseline he evidenced far less than about 3 episodes per day and on many days he was without incident. By the last 20 days of treatment he had become so amenable as to be allowed to move to a less restrictive learning-disabled classroom. His disruptive episodes were monitored by that teacher and verified by the psychologist, but seldom did he perform any inappropriate behavior during the last 20 days of treatment (see figure 10). Error variance on the time-series for this subject was 1.509 and level change was -2.30. This resulted in a highly significant t test value of -5.33 with a p of less than .001.
Fig. 9--The total number of daily "acting out" episodes for Experimental subject number 7 during the first 20 baseline days and the last 20 days of treatment.
Fig. 10—The total number of daily "acting out" episodes for Experimental subject number 8 during the first 20 baseline days and the last 20 days of treatment.
Experimental subject 9 evidenced 3 Emotional Indicators on his pretest and 2 on his posttest. This seems to be an adequate assessment of his inclination toward "acting out" behaviors when one compares this with his daily frequency of actual "acting out." On the 20 days of baseline he averaged a little over 3 episodes per day; this dropped to about 1 during the first 20 days of treatment, (see figure 11). During the last 2 weeks of treatment it was decided that his behavior was so well under control that he could be permitted to partake of regular classroom activities for about 1 hour a day. As with other subjects in these circumstances, his behavior was monitored during this time for "acting out" episodes. Typically he behaved quite well during this time and showed real interest in partaking in mainstream class activities. Frequencies of "acting out" behavior as analyzed via time-series show an error variance of 3.147 and a change in level of -2.56. The t test value for this data was -3.86 which was significant with a p of less than .001.

Experimental subject 10 had Bender Gestalt protocols with "acting out" Emotional Indicators of 3 on the pretest and 1 on the posttest. This again appears somewhat consistent with the actual rate of "acting out" in the classroom. She showed erratic tendencies toward high rates of noncompliance and talking out in baseline and showed some particularly good days during this time (see figure 12). By the
Fig. 11--The total number of daily "acting out" episodes for Experimental subject number 9 during the first 20 baseline days and the last 20 days of treatment.
Fig. 12--The total number of daily "acting out" episodes for Experimental subject number 10 during the first 20 baseline days and the last 20 days of treatment.
last 20 days of treatment her behavior was well within what
one might call the normal limits of classroom disturbance.
During this last phase she averaged between 0 and 1 "acting
out" episodes per day requiring correction by the teacher.
Error variance for the time-series was 4.644 and level change
was -6.88 which was again significant at a $p$ of less than
.001.

All experimental subjects had thus displayed a statisti-
cally significant decrease in the number of "acting out"
behaviors performed per day. Though some of these subjects
had indeed demonstrated classroom behaviors very reminiscent
of normal, well-behaved children by the end of treatment,
others were still performing at unacceptable high rates of
aggression, talking-out, out-of-seat behavior, and general
noncompliance. The variation among those in the group was
considerable, yet it appeared that some progress toward
behavioral remediation was made.

Control subject 11 had a long history of violent episodes
among peers. His classroom behavior exemplified a medium-to-
high level of "acting out" throughout the course of the
entire semester (see figure 13). His Bender protocol indi-
cated 3 "acting out" Emotional Indicators on pretest and
3 such indicators on posttest. Likewise, despite classroom
intervention and one-to-one therapy, his rate of "acting out"
throughout the semester remained constant. Error variance
Fig. 13--The total number of daily "acting out" episodes for Control subject number 11 during the first 20 days of baseline and the last 20 days of treatment.
for this time-series was 6.016 and change level was -.19. His \( t \) test value was -.26 which did not reach any level of statistical significance.

Control subject 12 demonstrated somewhat infrequent episodes of disruptive behavior, but on particular days he seemed uncontrollable. However, most of his "acting out" consisted of simple noncompliance. His rate of such behavior remained constant throughout the entire semester, vacillating between 5 and 0 episodes per day (see figure 14). Bender protocols indicated no change in "acting out" tendencies as he demonstrated 3 such Emotional Indicators on pretest and posttest. Error variance for this subject was found to be 5.853 with a change in level of .41. The \( t \) test value was .46 which was not statistically significant.

Control subject 13 had a long history of classroom disruption. Her irrelevant talking out behavior was the most conspicuous feature of her "acting out." Classroom management was unsuccessful in curbing the frequency of such episodes as the staff of this classroom seemed unable to coordinate the management program. One-to-one individual psychotherapy and contingency contracts also proved somewhat ineffective. Her daily rate of "acting out" averaged just over 3 corrected episodes per day and was maintained throughout the whole of the term (see figure 15). Error variance for the time-series was 3.217 with a change level of .23. The \( t \) test value of .32 was not significant. The Bender's protocols for this
Fig. 14--The total number of daily "acting out" episodes for Control subject number 12 during the first 20 days of baseline and the last 20 days of treatment.
Fig. 15--The total number of daily "acting out" episodes for Control subject number 13 during the first 20 days of baseline and the last 20 days of treatment.
subject also indicated no change in her inclination toward "acting out." She demonstrated 2 Emotional Indicators of that phenomena on both pretest and posttest.

Control subject 14 demonstrated some slight improvement in his "acting out" tendencies as verified by both his daily rate of such episodes and his Bender protocols. This seemed very conspicuously a function of a well-trained highly-motivated teacher who followed behavioral strategy directives with enthusiasm and precision. Bender protocols showed a drop of from 3 to 2 Emotional Indicators on pretest and posttest evaluation and a significant drop in frequency of overt "acting out" episodes. Baseline "acting out" behaviors averaged over 6 per day but dropped to between 2 and 3 per day by the last 20 days of treatment (see figure 16). Error variance was 7.298, and level change was specified at -3.16. The t test of -3.39 was significant with a p of less than .01.

Control subject 15 showed a definite drop in frequency of "acting out" behavior which coincided with a reduction in the number of Bender Emotional Indicators. She averaged around 4 episodes per day during the initial 20 baseline days but dropped to about half that frequency by the last 20 days of treatment (see figure 17). Teacher control seemed to be the critical factor in this subject's change. Error
Fig. 16—The total number of daily "acting out" episodes for Control subject number 14 during the first 20 days of baseline and the last 20 days of treatment.
Fig. 17—The total number of daily "acting out" episodes for Control subject number 15 during the first 20 days of baseline and the last 20 days of treatment.
variance was 3.768 with a change in level of -3.52. The $t$ test was -7.57 which was highly significant at $p$ of less than .001.

Control subject 16 also demonstrated a remarkable drop in his frequency of "acting out" and his Bender protocol reflected this transition. Frequencies dropped from almost 10 per day to almost 2 (see figure 18). During the same time period his Bender dropped from 2 Emotional Indicators to none. This subject's time-series had an error variance of 13.169 with a change in level of -7.08. This produced a $t$ test value of -5.07 which was highly significant with a $p$ of less than .001. Again, this emotional/behavioral transition can probably be attributed to teacher influence although this subject's Bender protocol probably overestimates the degree of his improvement when compared to his fluctuating "acting out" rate during the last 20 days of treatment.

Control subject 17 showed some improved behavior during the last 20 days of treatment by comparison to his baseline rate. This subject displayed an amiable temperament but tended to drift off-task erratically throughout the course of the entire semester. He did, however, reduce his frequency of such off-task, disruptive behavior as the semester progressed (see figure 19). Error variance for the time-series was 9.526 with a change in level of -3.42. This produced a $t$ test
Fig. 18—The total number of daily "acting out" episodes for Control subject number 16 during the first 20 days of baseline and the last 20 days of treatment.
Fig. 19 -- The total number of daily "acting out" episodes for Control subject number 19 during the first 20 days of baseline and the last 20 days of treatment.
value of -3.86, which was statistically significant at a $p$ of less than .001. He displayed 3 Emotional Indicators on Bender pretests and 2 on his follow-up protocol. This subject probably had greater potential for improvement than was demonstrated by either his Bender protocol or his frequency of "acting out." Classroom management was maintained at a very low level of proficiency in his homeroom. However, even this minimal intervention, as carried out by less-than-cooperative staff, seemed to show some beneficial effect.

Control subject 18 gave no indication of change throughout the course of the entire semester. Behavior management in his class consisted primarily of aversive verbal control. Praise for on-task behavior was a relatively low probability event. The teaching staff in this classroom seemed unwilling and disinterested in learning behavioral strategies. Accordingly there was no change in this subject's frequency of "acting out" (see figure 20). Error variance was 3.939 and change in level was only .77. The $t$ test value of 1.33 was not statistically significant. This subject's Bender protocol remained at 2 on pretest and posttest.

Control subject 19 showed some conspicuous improvements in his rate of "acting out." He averaged over 8 episodes a day during the baseline but dropped to almost 2 by the last 20 days (see figure 21). Teaching staff in this classroom made a conspicuous attempt to follow psychological
Fig. 20--The total number of daily "acting out" episodes for Control subject number 20 during the first 20 days of baseline and the last 20 days of treatment.
Fig. 21--The total number of daily "acting out" episodes for Control subject number 21 during the first 20 days of baseline and the last 20 days of treatment.
directives although the efficiency with which these directives were followed left much to be desired. Therefore, the significant improvement in the student's behavior may not be an accurate estimate of his potential for change. Error variance for this subject was calculated at 6.778 with a change in level of -5.46. This produced a $t$ test value of -6.50 which was highly significant at a $p$ of less than .001. Bender protocols likewise reflected significant change as he dropped from 3 Emotional Indicators on pretest to 1 on posttest.

Control subject 20 proved to be something of an enigma. Prior to baseline, he met all criteria for placement in either the control or experimental group. During baseline he performed a very low level of "acting out" and continued to maintain this level throughout the course of the semester. There was no significant change in his frequency of "acting out" because of this sustained low rate (see figure 22). In fact, most days on baseline and treatment show no indication of classroom disturbance. This subject's teacher had excellent behavior-management skills and followed psychological directives with apparent enthusiasm. However, as stated, change in rate is obscured because of such low baseline frequency. Error variance was 1.067 and level change was only -.34. This resulted in a $t$ test value of -.77 which was not statistically significant. Paradoxically this subject displayed 3
Fig. 22--The total number of daily "acting out" episodes for Control subject number 22 during the first 20 days of baseline and the last 20 days of treatment.
Emotional Indicators of "acting out" on both pretest and posttest. This observation is inconsistent with almost all other sources of data throughout this study and verifies the notion that Bender Gestalt protocols indicative of "acting out" inclinations and actual "acting out" behavior in the classroom are not entirely consistent phenomena. However, even though no absolute relationship exists between the Bender as a predictor of actual behavior, the relationship which has been demonstrated for most subjects seems somewhat significant.

In order to further illustrate a relationship between the likelihood of disruptive behavior being emitted in the classroom and the Bender Gestalt as a predictor of such "acting out," correlations between these two criteria have been performed pre- and postintervention. It will be recalled that a previously described study by Caudle (1981) indicated a Pearson Product-Moment Correlation of .425 between the number of Bender Gestalt Emotional Indicators of "acting out" and the actual placement of these students in time out for performing disruptive classroom behaviors. This was a large N sample and as such the correlation achieved was significant at a p of less than .001. In the current study a relatively small sample was used in order to maximize behavior-management impact. A pre-intervention correlation was performed on this sample and an r of .1852 was
obtained between the Bender and actual "acting out" behavior. This correlation is not statistically significant, but one must recall that the group under observation was essentially homogeneous; that is, all members of the control and experimental groups were performing high rates of "acting out" and had a relatively high number of Emotional Indicators on the Bender protocols. In general it can be said that the greater the variability among observations, the greater the likely value of $r$. The pretest Bender protocols and baseline "acting out" frequencies are an excellent example of restricted variability precisely because all subjects were chosen on the basis of their behavioral excesses. However after treatment, all of the experimental subjects and half of the control subjects had shown some improvement in their classroom behavior as well as their Bender protocols. When a correlation was run on the postintervention number of Bender Gestalt Emotional Indicators for each student and each student's average "acting out" frequency, it was found than an $r$ of .4565 was obtained. This correlation, even with such a small $N$, was found to be significant at a $p$ of less than .05. This corresponds rather closely to the results obtained in the Caudle study with an $r$ of .425. The intervention procedures had the obvious effect of diversifying the group's classroom behavior as well as their production of Emotional Indicators on the Bender protocols. This result further
corroborates the influence of behavioral strategies on disruptive classroom behavior in conjunction with Bender's indicators of "acting out."

Research hypothesis 5 specified a decrease in "acting out" behavior as a function of differential reinforcement of positive interactions among students and between staff and students and that this process would result in a decrease in the previously described Emotional Indicators on the Bender Gestalt. The previously described posttest correlations together with group and individual reductions in rates of "acting out" analyzed via time-series, plus improvements in the Bender protocols, all provide evidence which supports the acceptance of the fifth research hypothesis.

The analysis of covariance maintains four underlying assumptions regarding the nature of the observations under consideration: (1) that they be normally distributed, (2) that they have equal variances, (3) that they be independent, and (4) that the slope of the regression line be equal within each of the populations. The t test requires only the first three of these assumptions.

Although the analysis of covariance is robust with regard to the assumption of homogeneity of regression the possibility of a violation of this assumption was tested and no differences were found in the slope of the
regression lines. The analysis of covariance and the t test are both robust with regard to the assumption of homogeneity of variance, particularly when equal n's are used. However, the hypothesis of equal variance was tested via the F ratio and no significant differences were found between the pretest variances. In addition, the Central Limit Theorem indicates that regardless of the distribution of parent population means, the distribution of sample means becomes normal as n increases. Thus, even for non-normal populations, the sampling distribution rapidly approaches normality with n's as large as 20.

There are no equivocations regarding the assumption of independence. When observations are paired or pretest/posttest observations are used, statistics which take this process into account must be used. The analysis of covariance and the t test for dependent groups are such appropriate procedures. Thus, the previously described data adequately fulfill the underlying assumptions regarding proper distribution and quantification of the sample scores.

There has been some controversy regarding the propriety and necessity of statistical procedures when analyzing frequencies of single-subject data (Glass, 1975). In all single subject frequency data presented in this paper, significant differences in rate were conspicuous to the naked eye as well as on the time-series evaluation.
The question then arises as to whether the use of single-subject statistical procedures adds anything to the evaluation of such a design. Indeed they are essential since one cannot know, a priori, that changes in response rate will be visually apparent. Glass (1978) reports cases in which naked-eye observation of single-subject graphs depicting rate of response do not adequately confirm significant changes in frequency over time. In such nebulous cases, Glass indicates the need for probability-based procedures such as time-series. This is essentially the same argument used to defend large-group statistics. When means of large groups are conspicuously different, statistics appear superfluous. When the means are close, we want to know if the difference between such observations are real or a function of chance variation. It is the author's position that one cannot logically argue against the use of procedures such as time-series analysis without, at least by implication, arguing against the use of the more commonly used large group statistical methodology. Many behavioral psychologists have indicated that if changes in rate or means are not obvious by immediate visual inspection of the data, then, at least with regard to experimental analysis, the researcher has not adequately controlled competing variables or has not produced an experimental effect worthy of consideration.
Discussion

The study conducted by the author in conjunction with school district staff resulted in a number of specific determinations regarding the pervasive effects of contingency management in the classroom. The problem as stated in the introduction was to determine the degree to which behavior therapy and classroom management would be capable of affecting not only particular target behaviors, but also the more subtle and broad-range personality factors as measured by a popular index of emotional disturbance, the Bender Gestalt. The method for determining such an effect was by way of both individual and group statistics. The direct effect of contingency management was found by counting the frequency of disruptive episodes in the classroom. Such disruptive episodes, referred to as "acting out," were only counted when the teacher, psychologist, or teacher aid found the behavior severe enough to warrant a warning of forthcoming placement in time out or actual placement in time out. Time out was defined as placement in an isolated hallway for a period not to exceed 5 minutes or until the disruptive individual had become calm, for a period of at least 2 minutes.

The primary method of reducing the frequency of such "acting out" was through the reinforcement of behavior which was incompatible with disruption, i.e., appropriate social interaction and on-task behavior. Reinforcement was
given via tokens exchangable for primary reinforcers (assorted nonsweetened snacks), concurrent social reinforcement (verbal praise from the staffs) and free time given to play various games or individual recreational activities.

Reinforcement was given for successive approximation of such appropriate social interaction with peers and staff and for on-task behavior. A number of hypotheses were generated from this procedure which indicated the forthcoming beneficial effects to be derived by the subjects. The hypotheses were aimed at determining changes in explicit target behaviors by way of reducing frequencies of individual "acting out" behaviors and the more pervasive effects in terms of changes in group means of the Emotional Indicators of "acting out" on Bender Gestalt protocols. Individual changes in daily frequency of "acting out" were submitted to quantitative investigation and statistical significance for both experimental and control group frequencies while all individual members of each group were determined via the time-series experiment.

The research hypotheses for time-series analysis were stated and determined as follows.

1. It is hypothesized that there will be a significant decrease in the frequencies of "acting out" behaviors, over the period of the entire semester, for the contingency management group as a whole. This hypothesis was confirmed...
and changes in the experimental group's overall daily frequency of "acting out" were significant at a \( p \) of less than .001.

2. There will correspondingly be a series of significant differences in the "acting out" frequencies of each individual subject in the contingency management experimental group as measured over time. Research hypothesis 2 was confirmed in that all members of the experimental group showed individual significant decreases in the frequencies of their baseline "acting out" behaviors. Every member of the experimental group showed significance with a \( p \) of less than .001; however, at least two members of this experimental group still exhibited exceptionally high rates of "acting out" during the last 20 days of treatment. Thus, finding a significant decrease in frequency is not necessarily indicative of an overall behavioral remediation.

3. It is anticipated that there will be no significant differences in the frequencies of "acting out" behaviors of the control group as measured over time. This hypothesis was not confirmed as a significant decrease in the overall daily rate of "acting out," for the control group was established via time-series analysis. Various individuals within the control group were exposed to teachers who proved to be quite capable of following instructions regarding applied behavior analysis and thus their students showed definite improvement in their classroom behavior. The only real
difference that existed between the experimental and control group treatments was in the degree of expertise of the various teachers; it is therefore not surprising that about half of the control subjects showed improvement by a reduction in "acting out."

4. There will be a corresponding lack of significance in the "acting out" frequencies of individual members of the control group. This hypothesis was only partially confirmed as a great diversity of outcomes evolved in the control group. As previously stated, many of the eight control-group teachers showed a remarkable capacity for developing skills and following instruction in behavior-management techniques. It was not anticipated that as many as half the control subjects would show significant decreases in their frequencies of "acting out." Several of these control-group teachers showed a strong determination in developing the needed skills to manage these previously obstreperous students. This is perhaps one of the happiest findings in the whole study. It shows that a teacher need not have elaborate previous training in behavior management, as was true of the experimental group teachers, in order to have an impact on a child's tendency toward disruption. However, this finding further illustrates that those teachers who seemed least inclined to become involved in the process of learning behavior management skills had a minimal impact on their respective control students.
5. In a more general sense, it is hypothesized that a decrease in "acting out" behaviors, which occur as a function of differentially reinforcing increased positive interactions among students and between staff and students, will, because of common contingencies, eventually be reflected in a significant decrease in the number of indicators of emotional disturbance on the Bender Gestalt. This hypothesis was confirmed as an analysis of covariance on mean number of Emotional Indicators on the Bender Gestalt, on pretest and posttest for experimental and control subjects, a $t$ test on the pretest-posttest experimental group mean number of Emotional Indicators, and a Pearson Product Moment Correlation between the mean number of "acting out" episodes during the last 20 days and the number of Bender indicators on posttest, all showed statistical significance. These findings, together with the significant decreases in "acting out" episodes, all corroborate the confirmation of this hypothesis.

The hypothesis for group means showed the following results.

1. The null hypothesis ($H_0$) for the analysis of covariance states that there will be no posttest differences in the means of Emotional Indicators on the Bender Gestalt between the experimental contingency management group and the standard special education control group. This null hypothesis was rejected at a $p$ of less than .05. However,
various subjects in the control group who displayed improvements in their classroom behavior as a function of the contingency management procedures showed a corresponding improvement in their Bender protocols indicative of "acting out." Therefore changes in posttest scores between the experimental and control group would not have been as great had the control groups not had benefit of these procedures. Ethical considerations required the concurrent treatment of both groups and the differences which were found between groups are again probably attributable to the sophisticated behavioral skills of the experimental group teacher.

2. The research hypothesis ($H_1$) for the analysis of covariance states that there is a difference in the post-test means of Emotional Indicators between the two groups. This research hypothesis was accepted with the rejection of the previous contradictory null hypothesis.

3. The null hypothesis ($H_0$) for the $t$ test on the experimental group states there are no differences in the pre- and postintervention means of Emotional Indicators on the Bender Gestalt. This null hypothesis is rejected at a $p$ of less than .001 and is probably a better indication of the effect of the contingency-management treatment than the preceding analysis of covariance which was partially contaminated by the ethical requirement of concurrent treatment of the control group subjects.
4. The research hypothesis \( (H_0) \) for the \( t \) test on the experimental group states there will be a difference between pre- and posttest means of Emotional Indicators on the Bender. This research hypothesis is confirmed with the rejection of the preceding contradictory null hypothesis.

**General findings**

The use of two concurrently dependent but not quantitatively related variables is not a common practice in behavioral research. The findings of the current study indicate the Emotional Indicators of "acting out" behavior on the Bender Gestalt do, in fact, correspond rather well to changes in overt rates of "acting out" behavior.

Specifically this research shows a decrease in the frequency of "acting out" behaviors in the classroom analyzed via time-series, which are a function of reinforcing incompatible, on-task behaviors and appropriate social behaviors, has the concurrent effect of reducing the average number of Emotional Indicators of acting out on Bender Gestalt. The reduction of "acting out" indicators was determined by comparison to a control group of students, who did not all have the benefit of teachers with expertise in behavior management, and by experimental group comparison of pretest/posttest scores on the Bender via a \( t \) test. The analysis of covariance on the Bender indicators of "acting out" partial out the effect of any pretest differences in the experimental and control group and found a significant difference at a \( p \) of less than .05.
Finally, posttest "acting out" indicators on the Bender correlated at .456 with the average rate of "acting out" episodes occurring during the last 20 days of treatment.

These findings may be as important for the behaviorist as they are for the traditional psychologist. For the behaviorist such information gives an index of generalization to behaviors outside the classroom context. For the traditional psychologist, this information may be a more bitter pill. It has long been popularly misunderstood by the more traditionally-oriented psychologists that behavioral strategies are narrow in scope and that such treatments are not likely to have global effects (Sundberg, 1962).

This research indicates rather clearly that this need not be the case. In fact, the statistical tests show quite obviously that bringing the students under stimulus control in one context, the classroom, may well have the pervasive effect of influencing personality factors on a test of emotional disturbance. This outcome is not only a partial rebuttal of popular misconceptions regarding such phenomena as symptom substitution but it also serves as a challenge to other theoretical biases to establish the pervasive effects of their treatments. Will play therapists and other non-directive therapists replicate these results? Further research by such divergent schools of psychology is needed in order to establish the most efficacious, pervasive, and
parsimonious means of providing therapy for our emotionally disturbed children in the public schools.

Conclusions

This research has not only found a significant relationship between overt "acting out" behaviors in the classroom and Bender Gestalt indicators of the same phenomena, it has demonstrated a causative relationship between the two. There is, perhaps, nothing very new and different about reducing the frequency of disruptive behaviors of a class of so-called "emotionally disturbed" children. Behavior modifiers have had varying degrees of success with these procedures for at least the last 15 years. However, as stated in the review of related literature, behavior analysts have traditionally not been very much interested in outcome measures which are not directly related to changes in rate or form of target behaviors. Indices of personality as derived from "personality tests" have been seen as somewhat subordinate to the more basic goals of response change. But the notion of a more pervasive behavioral outcome should be of concern to those who seek information about the likelihood of response generalization. Behavior analysts could go beyond the immediate rate of classroom behavior when analyzing the significance of classroom management strategies. We might ask what effect we have had on the individual's overall propensity toward appropriate behavior in various settings outside the classroom. Bender "Emotional
Indicators" have been correlated with the child's general tendencies toward inappropriate behavior in diverse settings (Koppitz, 1963), and there is at least a growing body of research which indicates that "Emotional Indicators" on the Bender are well correlated with specific "acting out" tendencies in settings outside the school (Caudle, 1981).

The question arises as to why particular observations in the copying of the stimulus cards are so highly correlated with disruptive behaviors. What is there about producing dashes rather than circles, dashes instead of dots, uneven figures, progressive increases in size, sharp angles rather than wavy lines, etc. that relates to "acting out"? Any specific descriptions of an individual's motivation for producing such variations would be purely speculative. However, in a general sense, it can be assumed that many of these deviations from the stimulus card designs are somewhat easier to produce than the copying of the exact design formation as presented on the stimulus card. Thus an indifferent subject might be more likely to produce such aberrations. Certainly dashes are easier to draw than circles, sharp angles are more easily produced than smooth rounded angles, variations in size of designs, both large and small, occur when individuals show little concern for exact duplication of design size parameters; the same can be said of uneven figure size. Overwork is defined as impulsive, redrawn, and heavily reinforced lines. Many
of such inaccurate reproduction can be attributed to the subject's nonattending behavior and tendency to be easily distracted. Emotionally disturbed children typically have a history of noncompliance or at least indifference to following the directives of teachers and parents. It seems highly unlikely that any individual who is not disposed toward following the directives of a classroom teacher would be inclined to try to impress the examiner with exact replications of seemingly irrelevant designs. Rather, reproducing the design in a fashion which seems the least strenuous or perhaps, in other cases, reproducing the design in ways which are more interesting to the student are likely events in the testing of emotionally disturbed children. Both such strategies are reminiscent of behaviors which do not comply with the request of the examiner, i.e., to "copy the designs as you see them on the cards."

Children who are emotionally disturbed are typically not renowned for attending to instructional formats nor are they likely to comply with simple requests to behave appropriately. Classroom management is a procedure which produces a conspicuous structure within which obeying the teacher's instructions and following her directives is differentially reinforced while noncompliance and disruption result in verbal warnings or time out. Increased attending and complying behaviors may generalize to the Bender.
Interestingly, this interpretation is also consistent with Sprick's (1981) observation that disruptive children have poor "self-concepts" and that improvement of the child's self concept is related to decreases in the frequency of "acting out" episodes. The primary strategy employed throughout the semester was a tactic of reinforcing both on-task behavior and appropriate social interaction. As such, many of the students approximated increasing levels of such target behaviors and less inappropriate behavior was documented. Increased socialization and appropriate peer interaction together with increased academic skills all tended to improve what has frequently been termed "self-concept." Accordingly these changes were seen not only in the daily academic work of the students but in many approximations of written and oral work. Bender protocols which previous to treatment were done in a reckless, nonattending or self-serving fashion were, upon completion of treatment, produced in a more complying and controlled fashion. It is speculated that students who feel good about themselves might be more willing to try to impress the examiner with the accuracy of their reproductions than those who have a disregard for the tester and the testing format. The significant positive correlation between disruption and Bender protocols together with the significant changes in "acting out" frequencies in conjunction with the decreased
number of "acting out" indicators all seem to indicate that as these children came under the influence of the teacher, they strived to be more consistent with her requests; later they also made attempts to have their drawings coincide with the productions on the stimulus card.

In many ways this boils down to a simple matter of social control. Students who demonstrate social control are reinforced by the presentation of verbal praise statements. Those lacking in such social control typically are not influenced by praise statements of teachers or peers. Many may in fact work to counter control when praise statements are made contingent on appropriate behavior. An example of this phenomenon is easily seen when implementing a variation of the Long, Williams (1973) group contingency program. This program consists of giving the whole class free time for group-appropriate classroom behavior. However this program also provides for group loss of free time when individuals in the class are disruptive. Thus if any one individual should "act out" during class time, the whole class would lose a portion of free time. Concurrently as individuals behave appropriately the whole class gains in free time to be allocated at the end of the period. This can be a particularly strong form of social reinforcement. Such a technique has been an especially easy and effective classroom management strategy with normal and even mildly
disruptive students. But, as the student population becomes more disturbed, the probability of influencing class behavior with this technique drops drastically. As stated, the seriously disturbed "acting out" child apparently is not easily influenced by the praise of the teacher or the class. This was seen to be precisely the case with the children in the experimental classroom. It was only by pairing tokens (exchangeable for primary reinforcers) with praise statements that many of the experimental students began to come under the influence of verbal praise. It was not until almost the end of the term that the Long, Williams group contingency showed any positive effect with these children. Perhaps, in many ways, one definition of "emotional disturbance" relates to the extent to which an individual is capable of being reinforced by the praise statements of teachers and, perhaps as importantly, the "opinion" and praise of their peers.

All of the before-mentioned data and extrapolations from data provide a rationale for using the Bender Gestalt as a supplementary tool to direct data analysis of behavioral rate. While rate will always serve as the primary criterion for any change in overt behavior, instruments such as the Bender are beginning to show value as an index of generalization. It may become increasingly important for psychologists, teachers, and diagnosticians to make
some probability statements about the degree to which change in individual classroom performance will generalize to circumstances outside the class and to classes which are not highly structured. Though this study does not deal directly with the extent of Bender generalizations to other classes, this study does provide evidence for a cause and effect relationship between behaviors in the classroom and the Bender protocol. Further, several subjects in the experimental group who showed simultaneous reduction in emotional indicators and overt "acting out" were able to successfully maintain appropriate behaviors in other outside-class activities all or part of the day.

The primary utilitarian value of this study relates to the use of the Bender "acting out" indicators when rate has not been taken. For a multitude of reasons many of the files on children in special education contain Bender protocols but surprisingly few have any reliable data on actual "acting out" behavior. It is important to know how much confidence one can have in such an assessment device when more longitudinal data is not available. It appears from the results in this study that in fact the Bender may be one of the measures of personality that actually coincides with overt behaviors on a relatively consistent basis.

Further the Bender may begin to serve more and more as a bridge between theoretical biases. Both the traditionally
oriented psychologists and the behavior analyst may feel reasonably comfortable interpreting the Bender as an index of the same phenomena (at least with regard to "acting out" tendencies). In another sense this study serves as a point of information for many of the traditionally oriented psychologists who may react unfavorably to behavior therapy. Many traditionalists, like those described in the review of related literature, believe that changes in overt behavior bear little relevance to inner "personality" change. These therapists often seek to use procedures which aim to "cure" the underlying "causes" of psychic disorders. Everything from "free play" to "reflection" has been used to uncover the seeds of "psychic ailments" and to cure the "sick" personality. It may be noted that many if not most such therapists (Koppitz, 1973) accept the Bender as a key indicator of such a diseased "personality."

This study, however, shows quite clearly that design-drawing behavior can be brought under the influence of contingency management strategies used in the classroom. It shows that a commonly accepted index of "personality" may be influenced by a learning-theory approach. Although this may take some of the mysticism out of the Bender, it does not in any way debunk the test. Rather, this study points a finger at some of the variables for which disruptive classroom behavior and Bender protocols of "acting out" are
both a function, namely social control (i.e. the probability of coming under the influence of teacher verbal praise). The extent to which students are willing to comply with verbal instructions contingently related to social reinforcement seems to predict what has frequently been termed as emotional stability.

None of these transitions from disruptive to appropriate behavior were possible without the impact of the primary independent variables—the teacher and teacher aids. These variables quite clearly determined the degree of effect. For those students in this study who were exposed to the tactics and excellent behavioral skills of the experimental class teachers, some improvement was seen in their respective behavior and Bender protocols. For many of the children in control classes a comparable improvement was documented. Regrettably there were control students who demonstrated no significant change in behavior.

In all of the documented cases it is believed by the researcher that the primary determinant of positive outcome resided in the teacher's willingness and ability to perform behavioral intervention. Many of the control group teachers demonstrated a surprising capacity and enthusiasm for such strategies; correspondingly their students showed remarkable improvement. It is anticipated that this study may lay the groundwork for continued efforts among more teachers for
the learning and implementation of contingency management strategies in the classroom.

Recommendations

Findings reported in this study are perhaps somewhat inconsistent with traditional conceptions of behavior analysis. It is now for the schools to decide which approach to behavioral remediation they choose to emphasize. The current state of therapeutic intervention throughout most school districts is essentially eclectic. Emotionally disturbed children are exposed to all manner of professional therapeutic techniques which seek to change superficial and underlying causes of disruptive, obstreperous, noncomplying, aggressive, talking-out, withdrawn, and basic "acting out" behaviors. Many professionals believe that a nonstructured, nondirective approach is most suitable for addressing the "problems" "within" the emotionally disturbed child. But the very term "emotionally disturbed" is misleading. It tells the parent, teacher, administrator, and professional that it is not the child's overt behavior that is in question; rather it is the underlying emotions which are somehow "disturbed." Such terminology adds nothing to an understanding of the independent variables for which the inappropriate behavior is a function and postpones a more empirical analysis of such problems. The use of nonstructured, nondirective, traditional approaches which frequently provide therapy
outside and away from the classroom are not only misleading, but they are unlikely to produce behavior change within the classroom context. Moreover, they perpetuate an all too pervasive notion of "mental illness" which is addressed via circuitous and antiquated systems.

It is the responsibility of "health" care professionals to produce outcomes in what has been described as "emotional disturbance" which not only addresses the "underlying cause" of such behavior, but also the manifestation of that "disturbance." Over manifestations are by definition quantifiable, i.e., one can rather straightforwardly count the frequency of such behaviors. If we are to know that a person is indeed changed as a function of our treatment, then measuring the change by way of counting either an increase in approximations toward target behaviors or a decrease in frequency of inappropriate behaviors would seem a desirable practice. However, a cursory examination of most school district files of emotionally disturbed children will reveal little of such data collection. The practices and procedures of behavior analysis are still in their infancy and time work; therapeutic methods in the school system currently prevail. But as research gives way to practice, it is to be hoped that the advantages to an empirical therapeutic methodology will become apparent.

State departments of education should become increasingly aware of behavioral strategies and the advantages they provide
for ameliorating the problems associated with classroom disturbance and a multitude of behavior disorders. Further, the quantification inherent in behavioral methodology supplies the relevant data needed for accountability. The procedures used in behavioral methodology provide an ongoing therapeutic milieu throughout the course of the entire classroom day. A teacher who is skilled in such tactics need not be wholly dependent on the psychologist or other "health care" professional for alleviation of classroom problems. She, with the periodic evaluation and support of the applied behavior analyst, may become the critical factor in improving all behavior related to the classroom and generalizing far beyond its boundaries. The adoption of such a system not only establishes more immediate and quantifiable results, it, in the long run, is likely to save tax dollars uselessly spent on less efficacious strategies.

Most research into methods of improving student behavior end with a plea for more research into the area of remediation and therapy. This paper will prove no exception. But much of the research which can be done need not be accomplished by the "health care" professional alone. Every teacher who becomes skilled at the application of applied behavior management in the classroom is a likely candidate for producing valuable sources of new information concerning these strategies. It is now for these regular and special education teachers to take up the task of quantifying their
results in order to further the joint goals of education and psychology.
Appendix A

Procedure Followed in Administering Bender Gestalt Test

Students were administered the Bender Visual Motor Gestalt Test in a manner quite similar to the procedure recommended by Koppitz (1963, p. 15) with one modification. Instead of presenting each student two sheets of paper, size 8½ X 11 inches, only one sheet of paper was presented. Students were given additional paper only if they requested it. After rapport was established, the test was introduced by saying the equivalent of: "I have nine cards here with designs on them for you to copy. Here is the first one. Now go ahead and make one just like it." Most questions were answered by saying, "Make it look as much like the picture on the card as you can." Students were given a #2 pencil with an eraser to work with. No time limit was set for completion of the test. When a student had finished drawing a figure, the card with the design was removed and the next card was placed in front of the student, until completion of the nine cards.*

*Reprinted by permission from Caudle (1981).
Appendix B

Scoring Key for Bender Gestalt

Emotional Indicators

1. **Confused Order**: Figures are scattered arbitrarily on paper without logical sequence or order. It is not scored if any logical order can be discerned, for example, from top down and up again, from left to right or right to left. No penalty is given for placing Figure 8 at the top if there is no room left at the bottom or side of page.

2. **Wavy Line (Figures 1 & 2)**: Two or more abrupt changes in the direction of the line of dots (Figure 1) or circles (Figure 2). A continuous, gradual curve or rotation of line is not scored for this category. A change of direction must involve at least two consecutive dots or circles. A single dot or column of circles out of line is not scored. This indicator is scored only once regardless of whether one or both figures show this deviation.

3. **Small Size**: One or more designs are drawn half as large as the design on the stimulus card. Size of each figure is measured in both directions. When a design consists of two parts, both parts must be reduced in size.

4. **Fine Pencil Line**: Pencil line so thin that it requires effort to see the completed design.

5. **Dashes Substituted for Circles**: At least half of all circles on Figure 2 are replaced with dashes 1/16" or longer. Dots for circles are not scored.

6. **Progressive Increase in Size** (Figures 1, 2, & 3): Dots or circles increase progressively in size until the last ones are at least three times as large as the first ones. Scored only once regardless of whether it occurs on one or three figures.

7. **Large Size**: One or more designs are drawn one third larger in both directions than designs on stimulus card. When a design consists of two parts, both parts must be enlarged.
8. **Overwork or Reinforced Lines:** Total design or part of it is redrawn or reinforced with heavy, impulsive lines. The design may be first erased and then redrawn or it may be corrected without any erasures. Heavily drawn dots are included.

9. **Second Attempt:** Drawing of design or part of it is spontaneously abandoned before or after it has been completed and a new drawing is made. Scored when two distinct drawings are made of one design on two different locations. It is not scored when a drawing is erased and then redrawn on the same spot over the original drawing. It is scored when a drawing is erased and a second drawing is made on a different location on the paper.

10. **Expansion:** Two or more sheets of paper are used. Also scored if front and back of one sheet is used.

11. **Progressive Increase in Size from Figure A to Figure G:** An increment in the size of the drawings over at least six figures.

12. **Sharp Angles on Figure 6:** Sharp angles replace any of the smooth curves in Figure 6.

13. **Collision:** Any part of one design overlaps with part of another design.

14. **Dashes Substituted for Dots:** At least two of the dots in any of Figures 1, 3, or 5 are replaced by dashes 1/16" or longer.

15. **Circles Substituted for Dots:** Circles replace at least two dots in any of Figures 1, 3, or 5.

16. **Uneven Figure Size:** Both large (see No. 7) and small (see No. 3) figures appear on the paper.*

*Reprinted by permission from Caudle (1981).
References


Glass, G. V. Personal communication, October, 1978.

Glass, G. V. Personal communication, February, 1982.


Handler, L., & McIntosh, J. Predicting aggression and withdrawal in children with the draw-a-person and Bender Gestalt. *Journal of Personality Assessment, 1971, 35*, 331-337.


Spivack, G., & Swift, M. *Devereux elementary school behavior rotary scale manuhal.* Devon, Pa.: The Devereux Foundation, 1967.


Wenrich, W. W. *A primer of behavior modification.*


