THE USE OF THE BENDER GESTALT FOR EVALUATING
ORGANICITY IN ADULT MENTAL RETARDATES

by

RICHARD J. DESANTO, B.A., M.A.

A THESIS

IN

PSYCHOLOGY

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

MASTER OF ARTS

Approved

May, 1974
LIST OF TABLES

TABLE 1: Ebel's Rater Reliability Summary 61

TABLE 2: Means, Standard Deviations and t values for Three Bender Gestalt Trials 63

TABLE 3: Analysis of Variance Trials One and Three 65
CHAPTER I

INTRODUCTION

Purpose and Scope

While the Bender Gestalt Test is probably the most widely used psychological tool for assessing brain damage, experimental confirmation of such application has been largely limited to extreme cases (Billingslea, 1963; Nadler, Fink, Shantz, and Brink, 1959). Attempts to refine the instrument have most frequently taken the form of specialized scoring systems.

The best known system today is probably that of Koppitz, which is designed only for children (Koppitz, 1964). With adults, the Pascal and Suttell system (1951) and that of Peek and Quast (1951) have been the most common foci in the experimental literature (Billingslea, 1963). However, the adult systems have been even less successful in brain damage assessment than those with children (Koppitz, 1964). Thus, the need for a validated scoring system with adults remains a focal issue in contemporary Bender research.

One system affording some promise with adults in discriminating between organic patients and those with
psychiatric disorders was developed by Hain in 1964. Hain devised a weighted system in which fifteen classes of Bender distortions were given higher objective weightings because of their close association with organic patient productions in normative data. Two studies have compared the efficacy of the Hain system with that of more prevalent scoring conventions in the Peek and Quast approach (Mosher and Smith, 1965) and the Pascal and Suttell system (Kramer and Fenwick, 1966). In both studies the Hain system compared favorably, differentiating organic from non-organic adults at a statistically significant level.

Among subject populations with whom organic pathology is a crucial issue, the multiple syndromes of mental retardation are of central importance, yet such subjects are not represented in prior work with the Hain scoring system. Accordingly, one purpose of the present study is to evaluate the Hain scoring system on adults diagnosed as mentally retarded. The investigation will be concerned with whether the Bender Gestalt Test, scored according to Hain's norms, can discriminate organic from non-organic mentally retarded adults. The Bender has been shown to be successful in this discrimination (Bensberg, 1952; Feldman, 1953; Pacella, 1965), but has never been evaluated with adult retardates using Hain's scoring system.
A second objective of the present work relates to the examination of potential refinements in Bender administration procedures. Certainly, a visible proportion of Bender research has been directed toward this end (Billingslea, 1963). For example, Sullivan and Welch (1948) introduced the recall administration and Canter (1966) developed a background interference procedure to facilitate the elicitation of perceptual-motor processes distinguishing organic from non-organic subjects. The present study focuses upon an administrative refinement, proposed by Pacella (1965) for similar applications with the retarded. Pacella reported findings suggesting that multiple Bender administrations within a single session will produce practice-based gains for non-organic retardates while such a performance increment will not occur with subjects whose retardation involves an organic etiology. With frequent difficulty in replicating Bender research findings and a specific concern for the retarded, the present study is directed not only at Hain scoring system norms but re-examination of the Pacella hypothesis.

Overview of the Study

Two groups of twenty adult retardates each were selected on the basis of medical diagnosis and administered three successive trials on all Bender designs. The first
group included individuals whose retardation had been attributed to organic factors while the second group's retardation was classified under some non-organic etiology (e.g., cultural-familial). A single examiner administered all trials while four professionals scored the protocols on a blind basis. Initial protocols as well as gain scores between initial and terminal protocols were then examined to ascertain differential effects between organic and non-organic subjects.

**Method of Presentation**

The literature on the Bender Gestalt has become extensive over the last thirty years. For purposes of clarity, this literature will be discussed in the next chapter under five headings, ending with the category under which the present study would fall. The five categories or headings are:

1. The Development of the Bender Gestalt Test,
2. The Bender Gestalt in the Diagnosis of Psychiatric Patients,
3. The Bender Gestalt as a Personality Test,
4. The Bender Gestalt as a Test of Intelligence, and
5. The Bender Gestalt as a Test for Organic Brain Pathology.

In the succeeding chapter, the study proper will be presented with method, results, discussion, and conclusions. The study will be summarized in the final chapter and examined in a perspective relating it to other research.
CHAPTER II

A REVIEW OF THE LITERATURE

The Development of the Bender Gestalt Test

Since it was first developed in 1932, the Bender Gestalt Test has become one of the most widely used tools of psychological evaluation. Twenty years after Bender had published her original monograph (Bender, 1938), a survey of 185 hospitals and agencies determined this test was used by psychologists almost as much as the Rorschach, the Thematic Apperception Test, and the Draw-a-Person Test (Sundberg, 1961). Such widespread use is based essentially on two premises: first, that visual motor behavior is a developed skill; secondly, that there exists a "normal range" for reproducing the nine geometric designs which constitute this instrument. It is assumed that deviations from this normal range may be due to aberrations in intellectual capacity and function, brain tissue or chemistry, perceptual accuracy, motoric behavior, emotional stability, need gratification patterns or the use of defense mechanisms (Billingslea, 1963).

In Instructions for the use of the Visual Motor Gestalt Test (1946) Bender drew attention to this diagnostic multiplicity.
The Test has been used as a maturational test in visual motor gestalt function in children, to explore retardation, regression, loss of function and organic brain defects in both adults and children, and to explore personality deviations especially where there are regressive phenomena. In the latter instance it fills the role of a non-social, neutral, apparently innocuous test in a battery of personality tests (Bender, 1946, p. 3).

Bender selected the nine designs from a number of Gestalt designs originated by Wertheimer (1923). Wertheimer believed one had to understand behavior in terms of the totally functioning organism and placed emphasis upon total, integrated patterns of behavior rather than simple stimulus-response analysis. Bender continued this theoretical approach in applying these designs to developmental theory and clinical diagnosis. She simplified some of the figures she selected to accentuate particular Gestalt features (Tolor and Schulberg, 1963; Hutt, 1969). She labeled the first design "A" because originally she utilized it as a practice or introductory design; the others were numbered one through eight. "Only test figures A, 3, 7, and 8 closely resemble the designs used by Wertheimer. The others have been modified, usually to simplify them or to accentuate some basic Gestalt feature (Bender, 1963)."

Bender defined Gestalt functioning as a response of the whole organism to a constellation of stimuli. The
responses obtained were seen not simply as a series of parts making a whole, but as an intricate integration of the whole setting of the stimulus with the integrated state of the organism. "Integration occurs not by summation or subtraction or association, but by differentiation, or by increasing or decreasing the internal complexity of the pattern in its setting (Bender, 1938, p. 4)." In her later manual (1946) she gave a very succinct description of how this theory of functioning applied to design reproduction.

Gestalt function may be defined as that function of the integrated organism whereby it responds to a given constellation of stimuli as a whole, the response itself being a constellation of pattern or gestalt. Integration occurs by differentiation. The whole setting of the stimulus and the whole integrative state of the organism determines the pattern of response. Any pattern in any sensory field may be regarded as a potential stimulus. Any resulting pattern is a sensory motor pattern. Every sensory pattern has its background and orientation in relation to spatial gestalt function. A series of sensory motor experiences involves temporal patterning. Any deviation in the total organism will be reflected in the final sensory motor pattern in response to the given stimulus pattern (Bender, 1946, pp. 3 and 4).

Bender's original research applied this theory of perceptual motor functioning to the maturational process and clinical diagnosis. She was interested primarily in the development of gestalt functioning and intelligence in children, yet she viewed the test as a tool which could be used in the diagnosis of retardation, organic brain disorders, aphasias, the major psychoses, and
psychoneuroses.

Certainly one major innovation Bender brought to the use of these figures was the copying of the designs rather than just describing them. The examiner makes notes of any deviant behavior during the course of the test. The behavioral observations and reproductions then serve as a basis for clinical evaluation.

Despite the stimulating conclusions of Bender's original research, it has been difficult for many to accept that such a simple instrument could be an accurate tool for determining such a variety of processes as levels of intelligence and retardation, organic and functional disorders, varying degrees and kinds of psychopathology, and personality characteristics. Research into each of these variables abounds, but the results are mixed.

The Bender Gestalt in the Diagnosis of Psychiatric Patients

Most studies on this subject have had similar research designs. The test typically has been administered to one or more groups of experimental and control subjects whose diagnoses had been established independently of the Bender Gestalt. The protocols were then given to either professionals or non-professionals, or both, to see whether they could establish the diagnosis with this instrument alone. In some instances the interpreters
used a system of clinical inspection or one or more of the various scoring systems which have been devised. Bender used essentially a system of clinical inspection to evaluate the protocols of both children and adults. Both of her manuals (1938, 1946) included tables of responses for each year of age from three to eleven. A suggested mental age is obtained by comparing the protocols to these "standards."

Bender's "standards" formed the first crude scoring system devised for children. These gave impetus to the development of a number of more intricate and complicated systems for adults. Hutt devised the first formal method of scoring and interpretation for adults, but little is known about this early system due to the fact that research and results were never published. Billingslea (1948) was the first to publish an objective scoring system; however, it proved too clumsy because it contained 137 scoring items. Pascal and Suttell (1951) later determined 105 scoring factors; each factor was given a numerical value and was scored if present. This system became a standard and the catalyst of much research during the 1950's. Another important system has been that of Peek and Quast (1951); although this study was published, it evidently has been mostly unavailable (Billingslea, 1963), and thus it has had only a very minor influence.
Each of these scoring systems attempted to relate specific Bender deviations or indices to personality characteristics. For example, Hutt descriptively defined twenty-seven scoring factors. He then explained how certain combinations or "groupings" of these factors seemed to represent various psychodynamics or diagnostic characteristics. Billingslea's 1948 publication defined 38 factors to be scored using 137 indicators to decide their presence or absence on any of the nine designs. To determine whether his factors could distinguish psychopathology, Billingslea tested 100 neurotic adults against 50 judged to be normal. He was unable to validate his factors; moreover, he noted that Hutt's pattern for a psychoneurotic record was not present in most cases.

The Pascal and Suttell (1951) system established a pattern for developing other scoring methods. Like Bender, these authors gave the scorer actual records to assist in determining the presence or absence of one of their 105 factors in a particular protocol. Items scored as present yielded a total raw score for the subject. This score was weighted and interpreted on the basis of their standardization study. Their standardization norms included the productions of 271 adults with high school backgrounds and 203 college graduates. The raw scores of these two groups were translated into standard scores permitting the establishment of Z or weighted scores. To
test the diagnostic validity of this system, the authors
applied it to the protocols of three groups of psychotics,
neurotics, and normals, respectively. Significant
differences were found between the two groups with psycho-
pathology as well as between these two groups and the
normals. Later reproductive studies of similar designs
(Addington, 1952; Swenson and Pascal, 1953; Curnutt,
1953; Robinson, 1953; Lonstein, 1954) supported the
findings that the Pascal and Suttell scoring of the
Bender could discriminate groups of patients with varying
types of pathology from non-patient populations.

Other studies were less encouraging. Tamkin (1957)
was unable to discriminate psychotic from non-psychotic
male veterans using the Pascal and Suttell system. With
a similar design, Tucker and Spielberg (1958) were equally
unsuccessful in separating depressed patients from other
psychiatric disturbances among veterans. Earlier, Curnutt
and Lewis (1954) hypothesized that the Bender Z score
and the Rorschach F+ per cent of Beck were theoretically
similar, and should therefore be correlated; however,
their data showed no such correlation nor any discrimina-
tory value of the Pascal and Suttell system.

While these studies have apparently left the ques-
tion of the Pascal and Suttell system unresolved, they
illustrate a particular emphasis in Bender research at
that time. A broad spectrum of diagnoses was included in
all of these studies and many of the diagnostic groups were quite loosely defined. All followed the lead of Pascal and Suttell in attempting to differentiate psychotics from neurotics and these, in turn, from normals.

Using the Pascal and Suttell system, Curnutt (1953) successfully discriminated alcoholics from non-alcoholics, while Tucker and Speilberg (1956) failed to separate depressed patients from others. Addressing another issue, Mehlman and Vatovec (1956) submitted twenty-five protocols of matched patients to eight nationally known Bender experts. Five refused to participate in the study and the remaining three separated the groups at just slightly better than a chance level. It is apparent that high expectations were placed on this simple instrument. These early studies tacitly assume the operation of powerful effects, effects sufficiently marked to be manifest via either the single Z score index of the Pascal and Suttell system or the less formal clinical judgments of evaluation of varying levels of proficiency. Only two studies explicitly concern themselves with the diagnostic precision of clinical judgment versus objective Bender scoring.

Bowland and Deabler (1956) compared both clinical and objective scoring approaches. They submitted Bender protocols of four groups of similar adult males to both clinical inspection and the Pascal and Suttell system.
The groups were normals, neurotics, schizophrenics and organics. Their findings indicated that both approaches successfully separated the four groups, but neither method was superior to the other. Perhaps the most precise empirical address to this question has been that of Nadler, Fink, Shantz, and Brink (1959). First, they limited themselves only to the identification of organic brain pathology. Six judges tried to discriminate the protocols of twenty-six non-organic and twenty-seven organic patients. These six judges included two psychologists well-trained in the use of the Pascal and Suttell system, two psychologists who were not familiar with the system, and two occupational therapists not trained in psychology. Reliability between judges was high. The results indicated again that both clinical inspection and scoring were equally successful. Next, the twenty-eight protocols on the extremes of the distribution were removed; neither method could then discriminate the two groups significantly.

In the preceding literature, the Bender held up quite well when the problem involved the gross discrimination of protocols reflecting major deviations from normal behavior. Both the inspection system initiated by Bender and the objective scoring systems, especially that of Pascal and Suttell, performed equally well. Yet, when the task involved the discrimination of more subtle differences
in pathology between patient groups or the diagnosis of an individual case, neither approach was entirely adequate.

The Bender Gestalt as a Personality Test

The extent to which the Bender Gestalt is used by clinicians as a projective device revealing dynamics of personality is difficult to ascertain due to a limited number of empirical studies. Billingslea (1963) has pointed out that many clinical psychologists have produced these kinds of interpretations of Bender protocols and made them available to colleagues, but little published research has dealt with this usage.

Such interpretations are by no means an alteration of the Bender, for an impetus to Bender-based personality inferences was given by Bender herself. In the preface to the 1938 book, Schilder commented that "Dr. Bender does not forget that gestalt patterns are experiences of an individual who has problems and that the final configuration of experience is not merely a problem of perception but a problem of personality (Schilder, preface in Bender, 1938, p. ix)." Bender's position in this book regarding projective applications was more ambiguous. Her general position seemed to indicate little faith in the test's ability to delineate personality disturbances although she did speculate about its use in some functional
areas (schizophrenia, malingering, psychoneuroses, etc.). At one point she stated that "we do not expect to find disturbances in perception or in the visual motor gestalt function in the psychoneuroses," yet, she concluded the same paragraph on another note:

The unacceptable infantile demand for satisfaction is usually represented by some other activity which stands as a symbol for the real desires and drives of the personality. Since the stage of dawning consciousness is also the stage of maturation of the perceptual or perceptual motor gestalten, it would not be surprising to find that some such gestalten might become the symbol of the individual's unsatisfied infantile drives. In other words, they might represent the individual preoccupations, obsessions or compulsions (Bender, 1938, p. 157)."

After witnessing the projective use of the Bender Gestalt Test by some of her colleagues, Bender indicated grave apprehensions about the prevailing projective applications, noting how researchers and clinicians were misusing the Gestalt principles advanced in her original monograph (Bender, 1949). Later, she stated her opposition more emphatically. "Neither can ego strength be defined so as to correlate it with Gestalt function or other performances on the Visual Motor Gestalt Test. Since personality dynamics vary with different schools and sub-schools of psychology, the application of such dynamics to a perceptual motor test is spurious (Bender, 1963, p. xi)."

Bender's comments notwithstanding, use of the Bender Gestalt Test as a projective device continued unabated.

Suczek and Klopfer (1952) projected the nine designs
separately on a screen to a group of forty-eight beginning psychology students who were asked to write down free associations to each design. The authors were able to categorize the resulting associations under five headings: a list of objects frequently associated with each design, that part of the design upon which average interest focused, the affect pull of each design, the average symbolic meaning of the design, and the experimenters' tentative interpretation of the designs as symbols. Tolor (1957) duplicated this experiment with fifty Air Force psychiatric patients. He noted that the designs had decidedly different stimulus values for his subjects. Some of the designs showed many rejections, other elicited vague associations, and many were simply described. With such mixed results, he could only state that caution would be required when interpreting the symbolic import of Bender protocols.

Greenbaum (1955) asked children of what each design reminded them following their reproduction of all the figures. He mixed some of the nouns from each subject's statements into a commonly used word association list and later subjectively interpreted his subjects' responses to these word association lists.

Koppitz administered the Bender Gestalt and the Draw-A-Person tests to matched groups of sixteen first graders (1960). One group was studied under a nervous,
tension-producing teacher while the other group had a relaxed, non-tension producing mentor. The Bender protocols did not distinguish the two groups while the D.A.P. results were distinguishable. Gavales and Millon (1960) also used this stress, non-stress design with eighty college students. They first administered the Taylor Manifest Anxiety Scale and then divided the subjects into two groups of forty with high and low scores; half of each group was administered the Bender under artificial social stress while the remaining subjects were not. The stressed subjects consistently produced smaller figures regardless of their previous score on the anxiety scale; the same size relationship was found on their recall score. In a similar study, Lachmann, Bailey, and Berrick (1961) used the Taylor, Bender and the D.A.P. They noted that clinical psychologists' judgments for the presence or absence of anxiety in the Bender and D.A.P. were inconsistent and did not coincide with the scores on the Taylor Manifest Anxiety scale.

Other studies emphasized psychological factors other than anxiety and looked for positive indicants among Bender protocols. Wohl (1957) hypothesized that "constriction" should be generalized across an individual's Bender designs, Rorschach responses, Thematic Apperception Test stories, an interest test, and a semantic differential. Using an intercorrelational procedure, he failed to confirm
any generalization of constriction throughout these instruments. Prado, Peyman and Lacey (1960) used the Bender in an attempt to distinguish two groups of subjects selected on the basis of normal and "flattened" affect. The protocols were scored in a manner similar to Billingslea (1948) and again the two groups were not differentiated.

The most positive proponent of projective Bender applications has been Max Hutt. His approach will be examined in detail in the remainder of this section. Hutt presented a quite different approach to the projective use of the nine Bender designs. His method of administration was more elaborate than that commonly used. This more extensive administration was an integral part of the projective rationale which he proposed for this test.

He first pointed out that a variety of procedures has been developed for administration of the test. Many authors have departed from the original instructions of Bender, as she herself noted (1963), and a clearly stated rationale for each procedure using these nine designs has all too often not been given. Hutt hoped to avoid this pitfall. He specifically referred to his procedure as the "Hutt adaptation of the Bender Gestalt Test."

Hutt developed and used his own set of designs. In the early years following the publication of Bender's first monograph (1938), test cards were not available. Clinicians using the instrument often had available only
freehand reproductions or mimeographed copies which often contained irregularities in size, line quality, or inappropriate angles. Hutt decided to develop a set of "standard" designs in 1945 which he believed would be more closely representative of the original Wertheimer figures (Hutt, 1969).

An inspection of the currently existing Bender cards, published by the American Orthopsychiatric Association (1946), and Hutt's cards reveals three basic differences: size, line quality, and the nature of design number six. Hutt's cards are consistent in design size across all nine designs; the line thickness and darkness is smooth and unvarying, having the definite appearance of something printed. In contrast, Bender's cards still sometimes appear to be sketched. For example, the curved lines on design "A" (circle), design four (arch), and design six (sinusoidal curves) show the uneven thickness of a pressured pencil sketch. The dots on designs one, three, and five vary in size, having the appearance of being made by pressing a pencil point onto paper and twisting.

In Hutt's cards, design number six represents a major departure from Bender's original design which consists of two sinusoidal curves that intersect. Hutt's design number six has the Gestalt quality of being perceived as either two intersecting or two simply touching curves. The following sketches illustrate these differences.
Hutt developed these nine cards and his procedure of administration "specifically to maximize the projective features of this procedure so as to lead to meaningful predictions concerning personality characteristics and the possibility for their modification (Hutt, 1969, p. 14)." His method for administration entailed three distinct...
phases: the copy phase, the elaboration phase and the association phase.

The copy phase is essentially the procedure most clinicians regard as the standard administration. Hutt introduced some slight variations which should be noted. In the most common procedure, the examiner held the nine cards face down while giving the subject two medium soft pencils with erasers and a stack of 8-1/2 x 11 paper. The instructions were given (Hutt's were a little more elaborate), one sheet of paper was placed vertically in front of the subject, and card "A" was placed at the top.

In Hutt's version neither the position of the paper nor the number of sheets to be used was curtailed. The subject was free to change the position of the paper as he so desired. Hutt believed the subject's decision to rotate the paper was significant for interpretation. One explanation he advanced was the possibility of the existence of a tendency toward oppositional behavior. The subject's presentation of the paper in a vertical position and the card in a horizontal position was also seen to contribute to the occurrence of the phenomenon. In young children, this factor did, indeed, contribute to the frequency with which rotation occurred. (Hannah, 1958; Griffith and Taylor, 1961).

How the person structured the task was also deemed interpretably significant. Seeing that the examiner held
a number of cards in his hand, would the subject choose
to use one, two or any number of sheets of paper? Possibly
even one sheet per card? This, Hutt believed, allowed for
an idiosyncratic representation of the use of space and
for size variation. Although not strictly followed, other
examiners assumed that only one sheet would be used for
all nine designs. For example, Bender said: "It is well
to encourage the placing of the first figure near the upper
left hand corner of the paper although if the suggestion
is not readily accepted, it should not be insisted upon
(Bender, 1946, p. 6)." Hutt allowed for some degree of
freedom and ambiguity in the copy phase which was not
typical of other administrative procedures.

Hutt's procedure was unique in its use of an elaboration
and association phase. Following the copy phase, the subject's drawings were removed from sight. The subject
was then given a stack of paper and readministered the
cards with the instruction to feel free to modify or
change the drawings in any way he wished. The examiner then
used these productions for the association phase. Here,
each of the nine designs was placed alongside the subject's
respective elaboration drawing. The subject was asked to
give his associations to both the original card and his elaboration.

From this more extensive administration the examiner
had at his disposal more data for interpretation. This
included two samples of perceptual-motor productions (the copy and elaboration phases), the deviations between each, any sequencing deviations within or between the two samples, the verbal and non-verbal behavior noted and two sets of associations, one to the original design and another to the elaboration. Thus, it is immediately evident from this administration procedure that the "Hutt Adaptation of the Bender Gestalt Test" was specifically designed to "maximize the projective features of this procedure (Hutt, 1969, p. 14)."

The rationale for this projective use rested on the primarily non-verbal qualities of the task and, therefore, was not biased by the influence of language to the same extent as other projective tests. The Bender shared many common advantages with other projective tests such as the apparent capability to tap styles of adaptation, cognitive methods, areas of conflict, and specific defenses. Moreover, it offered additional advantages. Perceptual-motoric functioning that maturationally precedes language development could be assessed. Responses were less likely to be distorted by the subject. Conventionally appropriate or inappropriate behavior is typically mediated by language. In many projective tests, the subject can "filter out" inappropriate or revealing responses. On such a simple task as this, however, these conventions were less obvious so
the behavior tapped was not as readily distorted. Finally, the procedure had the advantage of offering additional data in those instances where verbal production was limited or non-existent.

Hutt's argument for the projective use of this procedure was probably best presented in the following statement.

The Hutt Adaptation of the Bender-Gestalt Test is an attempt to utilize this procedure as a projective device. It goes beyond the classical Gestalt laws of perception (namely, pregnanz, closure, nearness, and the like) and tries to understand both the process of responding and the final product in such ways as to maximize the understanding of the behaving individual: his idiosyncratic personality style; his needs, conflicts, and defenses; his level of maturation; and his coping methods and ego strengths. Such an approach can make use of both objective scores as well as complex clinical judgments. It utilizes the large body of evidence which has been acquired concerning the general nature of perceptual and motoric development, the nature of projective phenomena, and the effects of psychological and intracranial damage upon behavioral functions. It attempts to understand the individual's global functioning in the most parsimonious terms that will enable us to describe him and to predict some significant aspects of his behavior under defined circumstances. (Hutt, 1969, p. 2)

From the available literature it has been impossible to determine to what extent clinical psychologists may or may not be inclined to interpret the Bender projectively. Hutt stood alone in the development of and publication of this method. He certainly attempted to elicit a great deal more data than was usually assumed with the Bender and, for this reason, may be justified in the extent to which he carried his interpretations.
The Bender Gestalt as a Test of Intelligence

Psychologists are frequently requested to give a quick estimate of a person's intelligence. Unfortunately, most available instruments take considerable time to administer and score. Over the years considerable effort has been given to designing instruments which are less time-consuming without significant compromise in reliability and validity. With the increased utilization of the Bender Gestalt during the 1950's, its ability to serve this function was questioned.

Bender (1938, 1946) originally included a table of responses for each year of age from three through eleven. The subject's responses could be compared with these tables to obtain a suggested M. A. level.

Bender did not specifically define this M. A. as a measure of intelligence per se, but, rather, as a reflection of the maturational level of visual-motor perception or "gestalt function." She later pointed out that her studies indicated "that the visual-motor gestalt function is a fundamental function associated with language ability and closely associated with various functions of intelligence such as visual perception, manual motor ability, memory, temporal and spatial concepts, and organization or representation (Bender, 1938, p. 112)." Her studies also indicated the gestalt functioning necessary to reproduce her figures was usually fully developed by age twelve in the
normal population. With respect to intelligence, this latter conclusion has certainly been confirmed.

In later studies using "adults" (defined as twelve years of age and above) there appeared to be no significant relationship between the intelligence of adults and Bender performance. Of the few studies available, three have attempted to relate Bender scores with the Shipley-Hartford I.Q. Aaronson, Nelson and Holt (1953) and Peek and Olsen (1955) used adult subjects and found essentially no correlation between the Shipley-Hartford and a Bender recall score. Peek and Storms (1958) used three trained judges to estimate the intelligence of 100 hospitalized patients from their Bender protocols. They found no significant relationship between the judges' rankings and the subject's Shipley-Hartford score. Tolor (1956) correlated the usual recall scores with Weschler-Bellevue I scores on 175 patients and obtained an r = .50. Aaronson (1957) obtained recall scores and Porteus Maze I.Q.'s on 42 male and 46 female epileptics. The original correlation was .46 but, when age was partialed out, it was reduced to .21. This indicated some positive correlation with the children but not the adults. Thus, it seems reasonable to conclude that the Bender is not a good indicator of intelligence with subjects exceeding eleven years of age, but it does show promise with the 4-0 to 11-0 age range as indicated by the work of Koppitz.

Koppitz (1964) has developed the most extensively
and thoroughly researched scoring system for the Bender Gestalt. This scoring system was restricted to children and designed for two purposes: "to clarify objectively what level of performance can be expected from children at various ages... (and) to determine the significance of the different distortions and deviations on the Bender Test for children of different age levels (Koppitz, 1964, p. 4)."

Her procedure in designing the "Developmental Bender Scoring System" involved the initial selection of twenty carefully defined outstanding deviations. Only gross irregularities were selected since fine motor coordination is not highly developed in children. Distortion of shape, rotation, missing parts, confused order, overlapping, etc., were among the original twenty deviations.

Next, she administered the Bender to seventy-seven school children of average intelligence ranging in age from 6-4 through 10-8. Forty-one of these children were considered high achievers in school and well adjusted (good students) and thirty-six low achievers with poor over-all adjustment (poor students). There were forty-three first and second graders and thirty-four third and fourth graders. High and low achievers were equally distributed across grades. Achievement was selected as the validating criterion on the basis of a certain degree of maturity in visual-motor perception being assumed necessary to learn the
"three R's" in these early grades.

The twenty categories were then evaluated separately to determine which differentiated the good from the poor achiever. Seven of the original categories were found suitable for this purpose at a 5% level of significance. Each of these seven categories were then evaluated for their significance on each of the nine Bender designs. For example, "distortion of shape" appeared with significant frequency on designs "A", three, five and seven, but not on the other five. This analysis yielded thirty-one different scoring items to form the initial scoring system. These initial thirty-one items were again evaluated on the original seventy-seven subjects and a cross validation group of fifty-one patients at a guidance clinic. The results indicated that the composite score could significantly distinguish high and low achievers (Koppitz, 1958).

The system was later refined to include only thirty items and was again validated on 165 school children from six different schools. This time an item analysis was used against first and second grade achievement as measured by the Metropolitan Achievement Test. Finally, normative data for the Developmental Bender Scoring System was obtained from a public school population of over 1100 children in twelve different schools of various size and locale in Midwestern and Eastern states (Koppitz, 1960).

Since the Bender appeared to be correlated with
intelligence among children but not among adults beyond the level of maturation in visual-motor perception, Koppitz also did a study to evaluate the Bender and her scoring system in relation to intelligence.

Her subjects were 239 children between five and ten years of age. Each subject was first given a Bender followed by either the Stanford-Binet Form L or the Weschsler Intellligence Scale for Children (WISC) depending either on a chronological age of seven years or a suspected mental age lower than seven years. The subjects were divided into one group of 176 subjects with I.Q.'s of 75 to 149 (mean of 98) and another group of 63 subjects with I.Q.'s of 40 to 74 (mean of 63). Pearson correlations were computed between the I.Q. score and Bender score for each age level five through ten in the normal group and for all the subjects in the retarded group. All seven of these correlations were found to be significant at the one percent level. Koppitz concluded "the Bender Test can serve as a crude measure of intelligence for children age five to ten years. The highest single correlation was found at the five year old level which suggests that the Bender may be useful as a screening instrument for school beginners (Koppitz, 1964, p. 46)."

Two studies have confirmed the usefulness of the Bender as a screening device. Smith and Keogh (1962) administered both the Bender and the Lee-Clark Reading
Readiness Test to 149 kindergarten children and compared these results with reading achievement at the end of the first grade. Koppitz, Mardis and Stephens (1961) administered Benders and readiness tests at the beginning of the first grade and compared these scores with total achievement at the end of the school year. In both studies the correlations between Bender scores and readiness tests, and between these tests and the subsequent achievement tests, were significant at the .01 level or better.

So it appears that Bender's inclusion of a table of M. A. levels in her work was appropriate. The Bender Gestalt Test can serve as a crude measure of intelligence with children between the ages of five to ten years; however, once the subject matures beyond eleven years of age or his/her visual-motor gestalt functioning is fully matured, this instrument fails to differentially assess intellectual ability.

The Bender Gestalt as a Test for Organic Brain Pathology

A large portion of the studies devoted to the Bender Gestalt concerned themselves exclusively or in part with this instrument's ability to diagnose organic brain pathology. Koppitz (1964) estimated that fully one-fourth of the published studies dealt with this issue. This estimate would hold true today. These studies attempted to evaluate the Bender's ability to discriminate populations with organic pathology from those having no signs of organicity.
Essentially these studies are of two types: 1) those which attempted to discriminate organic psychiatric patients from other diagnoses of a functional nature (sometimes normals, or non-patients are also included) and 2) those which attempted to discriminate endogenous or familial retardation from exogenous or organic retardation.

In these studies the Bender procedure has often been varied in the hope of enhancing discrimination. In some instances, the results of the entire protocol were used with scores derived via the use of one of the scoring systems previously discussed. In other instances, specific signs or deviations associated with organicity (rotation, perseveration, distortion, etc.) were the main criteria. Finally, special administrations, especially the recall method, have been used to make the distinction.

Rotation of a Bender design has long been considered a distinct deviation resulting from brain damage. Yet, in those studies dealing with this factor, rotations are also found in the protocols of subjects with other diagnoses. Hanvik and Anderson (1950) compared rotations of 30 degrees or more between a group of forty-four brain damaged patients and non-patients. Over the nine Bender designs, 59% of the patients produced one or more rotations while only 19% of the controls did so. Hanvik (1953) examined the EEGs of twenty children having protocols with thirty degree or greater rotations and found sixteen of them
had abnormal EEGs. He concluded "the rotation of the figures of the Bender Gestalt test is even more highly predictive of brain damage in children than among adult patients (Hanvik, 1953, p. 399)."

Chorst, Spivack and Levine (1959) were suspicious of this conclusion because Hanvik neither used controls nor determined any percentage of abnormal EEGs among subjects not producing rotations. In reproducing Hanvik's study, they used sixty-eight "children" under the age of eighteen years with both an EEG and a scorable Bender. Similar to Hanvik, they found that fifty-one subjects had one or more rotations and that 69% of these had abnormal EEGs. Of the seventeen subjects not producing rotations, 47% also had abnormal EEGs. The probability of obtaining an abnormal EEG in their group was estimated to be 63%; thus, they concluded:

Using the presence or absence of one or more Bender rotations as a predictor of abnormal EEGs, we could make the correct decision in 65% of the cases. Using Bender rotations does not increase our predictive efficiency. Hanvik's (1953) conclusion that Bender rotations are highly diagnostic of brain damage in children is supported neither by his study nor by ours (Chorst, Spivack, and Levine, 1959, p. 559).

An earlier study by Fabian (1945) has also tended to throw suspicion on Hanvik's conclusion about rotation in children's protocols. Fabian found a high frequency of this deviation among kindergarten children. More than half of the very young children rotated the designs, but
only 7% of those who were between seven and one-half to nine years of age did so. Fabian also used modified horizontal lines and found that 51% of the six year olds and only 22% of the six and one-half year olds rotated the figures. Finally, he asked children to simply draw straight lines and discovered that almost 70% drew them vertically. He concluded that rotation toward the upright was a maturational process and that its persistence was a developmental lag or regression.

Hannah (1958) found that placement of the paper for copying was related to the frequency of rotations among children. When both the 8-1/2 x 11 sheet of paper and the stimulus card were presented in a horizontal position, the number of rotations was significantly reduced. Griffith and Taylor (1961) found similar results leading them to conclude that Bender rotations were, in part, due to a stimulus factor posed in administrative procedure.

Griffith and Taylor (1960) defined rotation as movement of a design forty-five degrees or more. In a population of 1,003 veterans hospital patients they found 56% of the mentally retarded and 40% of the chronic brain syndrome patients produced such rotations. These percentages were significantly higher than those in their schizophrenic reaction group, neurotic group, and character disorder group; yet, it should be noted that rotations were still present in protocols from each of these groups (23% for remaining
groups combined).

Other deviations have been suggested as indicative of brain damage but have not been singled out for independent study. Perseveration, or repetition of the whole figure or part thereof, is most often mentioned after rotation. It was considered significant by Barnes (1950), Bensberg (1952), Feldman (1953), Hain (1964), and Mosher and Smith (1965). Distortion of figures was deemed significant by Barnes (1950), Baroff (1957) and Beck (1959). Like rotation, however, these classes of Bender deviation do not appear exclusively in the protocols of brain injured subjects (Billingslea, 1963; Koppitz, 1964).

The use of a recall protocol in the identification of brain damage has often been mentioned. After the copy phase, the subject was asked to draw as many of the designs as possible from memory. This was not a routine procedure. It was probably introduced by Sullivan and Welch (1948) and gained some popularity through the work of Peek and Quast (1951).

The inconclusiveness of its value, however, may be one reason why only 20% of psychologist using the Bender have employed a recall protocol (Schulberg and Tolor, 1961). In addition to their rotation score, Hanvik and Anderson (1950) obtained a recall score on their groups of brain damaged subjects and control subjects by counting the number of designs correctly recalled. Mean recall scores
Niebuhr and Cohen (1956) compared four groups of subjects, 10 nurses, 10 acute schizophrenics, 10 chronic schizophrenics and 10 neurological cases, on the perceptual and motor aspects of Bender Gestalt performance. Each subject performed four successive tasks. First, the subjects were given a ten-second exposure to each of the nine designs. The exposure of each was followed by copying each design on a sheet of paper the same size as the stimulus card. Second, they were shown each of the nine designs and asked to copy them with no time limit imposed. Third, subjects were immediately asked to select each of the standard designs from six alternates from memory. Fourth, with the standard design before them, subjects were asked to select the matching alternate from the same six alternates. These procedures represent two reproduction tasks and two multiple choice tasks. The analysis of variance and intergroup comparison using a "t" test showed significant differences between all groups on the multiple choice tasks with two exceptions. Under task three, the memory condition, the difference between nurses and acute schizophrenics was not significant, while under the matching condition the
difference between the acute and chronic schizophrenics was not significant. In the reproduction tasks, difference between nurses and acute schizophrenics on the memory condition approached significance (p<.20), as did the difference between acute and chronic schizophrenics on the copy phase (p<.20). All other differences between groups on reproduction tasks were significant. The two memory conditions (tasks one and three) reflect a kind of Bender recall. On these two conditions Niebuhr and Cohen's results show significant differences between his ten neurological cases and his other three groups.

Tolor (1956) investigated the relative efficiency of the recall of digits (digit span) and Bender designs among ninety-one organics, thirty-five convulsives and forty-nine psychogenic patients. The three groups differed significantly in number of Bender designs recalled. The mean number of designs recalled by the respective groups were: organics 3.69, convulsives 5.5, psychogenics 5.53. Digit span differences were not significant. While the investigator found that Bender recall differentiated organics from psychogenics better than digits, it was not effective in predicting organicity in individual cases. Tolor later (1958) cross-validated this study using groups of schizophrenics, character disorders and organics. In this study, more precise controls for age and intelligence were incorporated by using a matching procedure. The
findings of this second study indicated that the organics had significantly poorer recall scores on both the Bender and digits forward, but not digits backward. The recall scores did not separate the functional groups.

Reznikoff and Olin (1957) conducted three recall studies for comparison. In the first study, they attempted to validate Tolor's initial work and selected their organics from Tolor's original population. The recall score was the number of whole designs reproduced. This score discriminated organics from schizophrenics at a .05 level of confidence. Unlike Tolor, however, they found no significance between the convulsives and non-convulsives. In two other studies, Reznikoff and Olin (1957) explored both the recall method and design difficulty in differentiating groups. Both studies used a modified scoring system adapted from that of Pascal and Suttell (1951). This score significantly discriminated both an organic brain damage adult patient group and schizophrenic adult patient group from a control group of normal subjects but failed to differentiate the two patient groups. Their analysis of scores for each design revealed that schizophrenics performed significantly better than organics with respect to design distortion and recall only on design four. Normals were significantly more accurate (.05 level) than schizophrenics only on design six, but they exhibited less distortion (.01 level) than organics on designs two, five, six,
seven, and eight.

Utilizing five diagnostic groups (organic, schizophrenic, depressive, neurotic, character disorder) and scores based on both the Reznikoff and Olin's (1957) scoring system as well as that of Hutt (1959), Armstrong (1965) compared the relative efficacy of scores based on the copy phase and the recall phase. She found that the copy phase differentiated the organics from both the schizophrenics and neurotics at a .01 level. The recall differences were even greater with virtually no overlap between organics and non-organics. Only 21% of the non-organic patients had as high or higher a recall score than the lowest obtained by an organic patient.

These studies have indicated that recall scores can differentiate organics from non-organics as groups. Armstrong's study suggested that, with refined scoring, better results may be achieved in individual cases. This is also implied in Stewart and Cunningham's (1958) use of a modified Pascal and Suttell recall score. The scoring system used does seem to refine the procedure somewhat. However, no one test factor, method of administration, or scoring system alone has been entirely adequate in diagnosing brain damage with the Bender Gestalt.

Another administrative procedure that demonstrated promise at least with brain damaged psychiatric patients
was the Background Interference Procedure (BIP) developed by Canter (1966). The procedure involved a standard administration followed by a second administration copied on a sheet of 8-1/2 x 11 paper described as a "confusing array of curved intersecting lines in contrast to the usual blank character (Canter, 1966, p. 91)." The subject's performance on the first administration was used as a standard of comparison with the BIP instead of some ideal or precise model. Canter's purpose in developing the BIP was to determine its efficacy in extending the sensitivity of the Bender Gestalt to separate deficits associated with organic brain disorders from those associated with other psychiatric disorders.

Canter (1966) first tested his procedure on a sample of eighty-four subjects divided into three groups, thirty brain damaged, twenty psychotic non-brain damaged and thirty-four non-psychotic non-brain damaged. Both the standard administration and the BIP administration were scored with the Pascal and Suttell scoring system. On the standard administration a significant difference (p<.01) was found between the brain damaged group and the other two. On the BIP administration the significant difference between the brain damaged subjects and the other groups (p<.001) was even greater. Distribution analysis between the two administrations also showed the BIP superior. On the standard administration over one-third of the brain damaged
group overlapped with the other two groups in the lower scoring ranges. On the BIP there was virtually no overlap. Only one of the thirty brain damaged subjects achieved a comparable score to subjects in the other groups. The remaining twenty-nine had higher scores. Scoring criteria were derived from analysis of comparative difficulty of design items and the total score on both administrations. These criteria were cross-validated on a second sample of sixty-five subjects with similar results. Twenty-one of twenty-six brain damaged subjects were correctly classified via the BIP score.

Later, Canter (1968) tested the reliability of his procedure's scoring system and further validated the BIP. Canter applied his procedure to two samples of forty and forty-two subjects referred for psychological evaluation. Scorers included an experienced clinical psychologist and a non-clinical psychometrist. On the first sample, scorers agreed on thirty-six of forty cases (r = .89). Thirteen of fifteen brain damaged subjects were correctly diagnosed while three of twenty-five non-brain damaged were incorrectly classified. On the second sample, twelve of fifteen brain damaged subjects were correctly classified and only one non-brain damaged subject was incorrectly categorized.

A student of Canter's (Adams, 1968) was less successful in applying the BIP to retarded children. Testing
thirty brain damaged and thirty non-brain damaged retarded children, he concluded that "the accuracy of prediction was too low to allow confidence in diagnosing the individual case. There was no evidence of a reliable difference in the efficacy of the two test forms (Adams, 1968, p. 2198)." Song and Song (1969) replicated Canter's original design on adult retardates with brain damaged (N = 33) and non-brain damaged diagnoses (N = 22). In addition, a third group of emotionally disturbed retardates (N = 15) were tested in an attempt to demonstrate that variations were not due to emotional factors. All three groups showed decrements in their scores from the standard administration to the BIP. Distribution overlap yielded no significant difference between groups. The range of BIP scores for the two non-brain damaged groups was between +20 and -20. Twenty-five of the thirty-three brain damaged subjects (76%) also scored in this range. While Canter's Background Interference Procedure demonstrated great promise with psychiatric patients, its validity with retarded subjects is clearly suspect.

One administrative procedure that demonstrated some potential with adult mental retardates was that of Pacella (1965). He also used successive administrations, but without any procedural or reproductive changes. Pacella tested twenty-two brain damaged and twenty-two non-brain damaged retardates, ranging in age from twelve to thirty-five years, on three Bender trials. He predicted that the latter group
would improve due to practice benefits while brain damaged retardates' performance would be unchanged. An analysis of variance on the first and third trials yielded no significant difference between groups on the first with significance at the .05 level on the third. He concluded that "for some mental retardates, visual motor abnormality is a learning deficiency for which they may compensate through repetition, while for the other mental retardates, it is the expression of some type of learning impairment which is not easily overcome by learning (Pacella, 1965, p. 727)."

Two studies which have attempted to bring together most of the Bender factors used in the analysis of organicy are worthy of more thorough discussion. Koppitz (1962) attempted to define the many complex issues of this diagnosis among perceptually developing children. Hain (1964) attempted to design a weighted scoring system to assist in the differential diagnosis of adults.

Prior to Koppitz, only three studies had concerned themselves with this diagnosis among children alone: Hanvik (1953), Shaw and Cruickshank (1956), and Wewtzer (1959). Koppitz's intention was to determine whether the Bender as a whole or in part could differentiate brain injured from non-brain injured children and whether the results would vary depending on age and intelligence. Her subjects were 384 school children from ages five through ten
years, divided into 103 brain damaged and 281 normal controls. The I.Q. scores of the brain damaged group ranged from 75 to 122 with a mean of 90 while no I.Q. was available for the controls. Controls were matched with the brain injured on age, sex and grade. All protocols were scored according to the Developmental Bender Scoring System for Children. These scores were analysed through the total score, each scoring item, and the relationship between I.Q. and Bender score for the experimental group. Chi-square values were computed for each matched age group of brain injured and controls. The results for each level were all significant at the .001 level. None of the five, six and ten year olds, and only nine subjects among the seven, eight, and nine year olds of the brain injured group had good Bender records when compared with normative Bender scores for their respective age level. On closer examination, it was found that most of the nine subjects with good Bender records had difficulty primarily in auditory rather than visual-motor perception and were able to compensate. This suggests that a good Bender does not, of itself, rule out brain damage.

On the other hand, a poor Bender score does not necessarily indicate brain damage. Among the controls, sixty-seven subjects (about 25%) also had poor Bender scores. Only a few such cases could logically be attributed to undiagnosed brain damage. Other factors such
as poor motivation, fatigue, illness, genetically induced impairment, or, especially, visual-motor immaturity represent the most likely determinants of this rather large percentage of poor Bender records among control subjects.

A poor Bender record may be thought of as indicating the possibility of brain injury especially if the Bender score is more than minus one standard deviation from the mean normative Bender score for a given age group. But a definite diagnosis of brain injury should never be made solely on the basis of a single Bender Test score . . . nor can the presence of neurological impairment be definitely ruled out because of a good Bender Test performance (Koppitz, 1964, p. 76).

Koppitz also computed chi-square values for each of the thirty scoring items marked as present in the protocols of her subjects. In general, all thirty items were scored in both groups and none appeared exclusively in brain damaged subjects. As age increased, the diagnostic significance of particular deviations also increased. Among the five and six year olds only a few specific scoring items were of value in diagnosing brain injury. A sampling of these computations indicated that rotation, for example, was not a significant deviation on any design until at least age nine. The rotations of designs four, seven, and eight had some significance at earlier ages when present with other deviations. Perseveration on designs one, two and six appeared frequently among all children through age seven. By age eight, perseveration appeared only on the protocols of brain injured subjects, rendering
this feature highly significant at subsequent ages. Integration of designs two and three, a rather gross distortion almost destroying the basic gestalt, was highly significant by age six. Finally, the substitution of straight lines for curves (design six) or lines for dots (designs three and five), though rare, was almost exclusively found among brain injured at all age levels.

Finally, Koppitz determined that the relationship between I.Q. score and Bender score among the brain damaged was not very close. The I.Q.'s of the 103 subjects were proportionally balanced with similar proportions of high average, average and low average subjects at each age level. Since 94 of these 103 subjects had below average Bender scores, I.Q. does not relate well with a good Bender score. Instead, "brain injured children as a group tend to do poorly on the Bender Test regardless of their I.Q. score (Koppitz, 1964, p. 83)."

In using the Koppitz system to evaluate a Bender protocol, one is provided with essentially two means of determining the diagnosis. A total score analysis can be used. By comparing the subject's total score with the normative data for that particular age group, an examiner can obtain a gross estimate of his subject's performance. The examiner can also perform an item analysis. By taking each scoring item present in the protocol, he can determine whether or not each is significant for his
subject's age group. From such a composite picture an examiner can, with some assurance, raise the probability of detecting the presence of brain damage.

The assessment of brain damage in adults is more difficult. As the reviewed studies have indicated, the Bender shows reasonable promise in distinguishing patient populations with organic pathology from non-organic normals but is less successful in differentiating such individuals from other patient populations. With this problem in mind, Hain (1964) devised a weighted scoring system to assist in the separation out of organic cases from others. His scoring system is aimed primarily at diagnosing brain damage in adult patients.

Previous studies have tried to use the Bender to distinguish organicity from schizophrenias, the different psychoses from each other, and these, in turn, from the psychoneuroses. Hain was less ambitious. In his original study he included only three groups: brain damaged, psychiatric cases, and normals. The brain damaged group (N = 20) included such diagnoses as cerebrovascular disease, and cerebral atrophy. The psychiatric group (N = 38) included diagnoses of varying degrees of severity such as depressive, obsessive-compulsive reactions, schizophrenias (paranoid, acute, schizo-affective, chronic), and psychophysiological reactions (tension headaches, torticollis, etc.) The normal group (N = 25) included staff
personnel and four patients judged to be without psychiatric disorders. From an earlier pilot study, Hain had selected thirty-one signs which seemed to discriminate between brain damaged and non-brain damaged states. He applied these thirty-one signs to the study's eighty-three subjects to determine the frequency of each sign's presence over all and between groups. Statistical weights were then given to a final fifteen of these signs. The highest weight was given to those signs which maximally discriminated the brain damaged subjects from the other groups. If a sign discriminated between psychiatric and control subjects, it was either given a lower weight or dropped in order to diminish the contribution of signs associated with psychiatric states.

This scoring system of fifteen weighted signs was first applied to the protocols of the eighty-three subjects. The difference between the scores of the brain damaged group and both of the other groups was significant at the .01 level. The difference between the psychiatric group and controls was not significant as expected. This lack of difference allowed for these two groups to be designated as one in order to determine the best single cut-off score between brain damaged and others. The best cut-off score of eight or below misclassified seven of the twenty brain damaged subjects, one from the psychiatric group, and one from the normal group. Because
of these seven false negatives and two false positives, and because the system had been developed in this group, a cross-validation study was performed.

The second study (Hain, 1964) was made up of a brain damaged group (N = 21) matched in age and intelligence with a psychiatric group (N = 21). The psychiatric group was used as a control in this study because in most clinical situations the goal is to differentiate between brain damage and other psychiatric disorders rather than between brain damage and normality.

The results of this study were consistent with the first. The difference between the mean scores of the two groups was significant at the .01 level. Again, however, the establishment of a cut-off score produced several false negatives (brain damaged not discriminated due to low score). Upon closer examination, Hain discovered that in both studies the Bender was best in diagnosing impairment of a diffuse nature such as that in arteriosclerosis. It was less successful with localized damage such as lesions and seizures.

In most of the preceding studies, the type of brain damage or its specific location in the brain was seldom clarified. Hain was one of the few to draw attention to this fact. The generally unspecified nature of the diagnosis "brain damaged" has lead some to suspect the diagnostic capability of the Bender Gestalt. Garron
and Cheifetz (1965), for example, have hypothesized that the Bender taps primarily parietal region function, and, indirectly, that of the temporal regions with the copy phase more closely related to the former functions and the recall phase to the latter.

While the Bender shows a good degree of success in discriminating groups of brain damaged from non-brain damaged, it has not been totally precise in correctly labeling each and every case. Moreover, many cases used in these studies were quite severe and entailed a less difficult diagnostic judgment. When the individual case is in question, deciding whether or not brain damage is present requires a great deal of caution. While the Bender may be suggestive of brain damage, it cannot always rule it out.
CHAPTER III

METHOD AND PROCEDURE

The Rationale for this Study

Previous studies have shown some promise in using the Bender Gestalt to detect brain damage in adult mental retardates. Bensberg (1952), Feldman (1953), and Baroff (1957) each took relatively large samples of matched brain injured (exogenous, acquired type) and non-brain injured (endogenous, familial, inherited type) adult retardates and compared their performances on the Bender designs. These studies indicated that the reproductions of the brain injured were inferior to those of the non-brain injured. Moreover, the correlation between Bender score and Mental Age was higher in the non-brain injured group than in the brain injured.

Various scoring techniques have been used to evaluate Bender performance and aid in the diagnostic discrimination of brain damage from non-brain damage. Bensberg rated his protocols according to Bender's (1938, 1946) portrayal of the average reproductions of normal children of different chronological ages. Temmer (1965) used a simplified system of her own design including ten types of
errors divided into errors of sequence, errors of reproduction and errors of relationship. Feldman (1953), Baroff (1957), and Pacella (1965) used the scoring system of Pascal and Suttell (1951) to establish their diagnoses.

The Pascal and Suttell scoring system is the most widely used in all Bender Gestalt research. While it demonstrates some success in discriminating the brain injured adult retardates in the studies cited, it should be noted that the scoring system was not designed specifically for this purpose. Instead, it was designed and validated for the differential diagnosis of hospitalized mental patients. Also, the 105 scoring items have proven cumbersome (Billingslea, 1963) and emphasis was not placed on an etiology of an organic or functional nature.

Hain (1964) designed a scoring system aimed specifically at diagnosing brain damage in adult patients using the Bender by assigning a greater weight to those deviations that appear most frequently in the protocols of brain damaged adults. He decided on fifteen scoring items with values ranging from one to four depending on frequency. Those deviations appearing most often in brain damaged protocols were assigned a value of four, and so on in descending order in comparison with psychiatric and control groups. Hain's system was the only Bender scoring system for adults or children which gave weight to specific deviations instead of simply scoring items present or absent.
Follow-up studies using brain damaged and psychiatric groups have compared both the Peek and Quast (1951) and Pascal and Suttell (1951) scoring systems with the Hain system. Mosher and Smith (1965) compared Hain's system with that of Peek and Quast on 142 brain damaged and 120 control patients. They found that the mean score from both systems could successfully discriminate at the .001 level of confidence. On the other hand, the suggested cut-off scores for both systems falsely classified over 70% of the brain damaged as being normal. They concluded that each system may aid in determining the presence of brain damage but cannot rule it out. Kramer and Fenwick (1966) compared Hain with Pascal and Suttell and also found that both systems successfully discriminated their groups at the .001 level, but Hain's system, because it was specifically designed for this purpose, was more successful.

The present study is intended to evaluate the use of the Hain scoring system with the Bender protocols of adult mental retardates. Hain's system was designed for diagnosing organicity in adult mental patients. No such system exists for evaluating this factor in using the Bender Gestalt with adult mental retardates. Koppitz's (1964) Developmental Scoring System has proven useful in evaluating organicity in both average and retarded children.
But once sufficient sensory motor maturation can be expected in adult retardates, it is difficult to determine the significance of Bender deviations with respect to the presence or absence of brain damage.

A secondary purpose of the present study was to evaluate the hypothesis and results of Pacella's (1965) study. He compared organic and non-organic mental retardates on three successive trials of the Bender Gestalt. He hypothesized no significant difference between groups on the first trial but superior non-organic performance on the third trial on the basis of practice effects which would not be reflected in organic subject performance. He believed learning would occur due to the repetition involved. "Organic patients are cerebrally deficient in coordinating perceptual experiences with motor behavior, and are therefore less likely to profit from learning, while non-organic retardates are simply slow to learn and will improve with repetition (Pacella, 1965, p. 724)." Using this successive method of administration, he hoped to improve prediction of the respective diagnostic classification.

Accuracy of reproduction over three trials and reproduction time were the two variables used. His results indicated that the average time to reproduce the three trials was no better than chance in classifying the brain damaged group. The percentage of correct classification
on the basis of speed alone for both groups was far less
than accuracy or improvement over trials. For these
reasons time has been excluded as a variable from the
present study.

Pacella's analysis of variance indicated no signifi-
cant difference between groups in accuracy for the first
two trials. On trial three he obtained a difference signi-
cificant only at the .05 level of confidence. Considering
just his first trial as a standard Bender administration,
Pacella's results show both some consistency and some
inconsistency with previous studies.

His accuracy of reproduction is based on the total
score achieved using the Pascal and Suttell scoring system.
Feldman (1953) also used this system to score his Benders
on a much larger sample of adult retardates (two groups
of fifty-four each, compared with Pacella's twenty-two
each). Feldman did find a difference between groups signi-
cificant at less than the .01 level. Assuming their popula-
tions were comparable, the respective results could be
considered inconsistent; one study achieved a significant
difference between Pascal and Suttell means and the other
did not. Yet, a closer examination of Feldman's results
illustrates a consistency, too. The range of scores from
the Pascal and Suttell scoring system were comparable for
both of his groups. Feldman's fifty-four exogenous
subjects had scores ranging from 21 to 271 (mean = 113.15),
while his fifty-four endogenous subjects had scores ranging from 26 to 214 (mean = 76.91). Thus, there was considerable overlap in performance accuracy between groups with only the means being significantly different while the ranges were comparable. With a score of 101 or higher, the chances were three to one that the subject was exogenous. Whereas, with a score of 60 or lower, the chances were three to one that the subject was endogenous. In the middle range between 81 and 120, prediction was about fifty/fifty. So, while both of these studies differ in Pascal and Suttell group means, neither demonstrates precise diagnostic reliability using this scoring system.

Baroff (1957) also used the Pascal and Suttell system to duplicate Feldman's results and to compare the Pascal and Suttell score with the Mental Age of an endogenous group of seventy-six retardates. Both studies indicated a high correlation between M.A. and Bender score, especially for the endogenous adult retardate. These results indicate the necessity of using subjects of comparable I.Q. or M.A. Baroff also tested 8 exogenous retardates and found that six of them had poorer performances (using the Pascal and Suttell total score) than endogenous retardates of comparable Mental Age. While the exogenous performance was inferior, the mean difference for these small comparison groups was, again, not significant.
In these studies the Pascal and Suttell scoring system has shown very limited success in discriminating brain damage from non-brain damage. Two factors may account for this. First, the Pascal and Suttell system was designed for differential diagnosis of a wide range of psychiatric patients using the Bender. Second, only nine deviations in the system were considered an "organic indicator" and these, like all other deviations, were only scored as present or absent. Several of these deviations were scored as present in a rather high percentage of cases in both Feldman's and Baroff's non-brain damaged, endogenous samples. For example, deviation number one, perseveration in designs one and two, were scored present in 39% of the endogenous cases in Feldman's sample and 34% of the cases in Baroff's.

By contrast, Hain's scoring system gave a greater weight to deviations which appeared more frequently or almost exclusively in brain damaged protocols. While both scoring systems were validated on psychiatric patients, Hain's intent was to aid in establishing this one diagnosis as distinct from all others. Both systems compare favorably in establishing this diagnosis among psychiatric patients with Hain's system slightly better (Kramer and Fenwick, 1966). For this reason, and the fact that the Hain system has not been used with adult retardates, it was selected for use here.
The Hain system will be evaluated in essentially two ways. First, it will be evaluated for interscorer reliability. All protocols will be scored separately by a matched group of professionals to determine whether the scoring norms can be applied objectively. Second, it will be evaluated as to its ability to diagnose brain damage. Since one administration of the Bender Gestalt is standard clinical procedure, this evaluation will be made on the first trial protocols. Furthermore, the Pacella hypothesis which states that there will be no significant difference on this and the second trial, will be evaluated using this system.

Procedures

Subjects were selected from the resident population of the Lubbock State School, Lubbock, Texas. Two groups of twenty subjects each with either a diagnosis of brain damage or no signs of brain pathology were gathered for comparison. Subjects were determined on the basis of chronological age, I.Q. and a medical diagnosis, including at least a standard neurological examination.

"Adult mental retardate" for purposes of using the Bender Gestalt was defined as a chronological age of 15 or older. A cut-off age of 15 was selected in order to rule out, in so far as possible, maturational factors in sensory motor development. In a normal population,
this has been established as eleven years of age on the Bender. A leeway of three years was allowed to account for the retardation and to safely assume that the maximum possible maturation of the subject for the performance of this task had been achieved. The average age of the brain damage group was 20.6 (median age 19), while the average age for the non-brain damaged group was 19.2 (median age 16).

The I.Q. of each subject was based on the most recent full scale I.Q. achieved on either the WISC, WAIS or, in four cases, the Stanford Binet Form L-M. In these four cases, a Weschler Scale was not available due to the low I.Q. These four subjects account for the lowest I.Q.'s in each group (brain damaged group, I.Q. 27, 44; non-brain damaged group, I.Q. 40, 42). The I.Q.'s for both groups are unusually close and therefore matched. The I.Q.'s for the brain damaged group ranged from 27 to 73 with a mean I.Q. of 56.5 (median = 57). Those for the non-brain damaged group ranged from 40 to 69 and have exactly the same mean I.Q. of 56.5 (median 56.5).

The diagnosis for the forty subjects was based on a medical examination which included a standard neurological examination exclusive of any psychological tests. Diagnoses for the twenty brain damaged subjects (twelve males, eight females) included: brain damage with convulsive disorder (six subjects), postnatal cerebral
infection (five subjects: three pneumonia, one tuberculosis, one measles), cerebral birth trauma, cyanotic, left-superior frontal involvement (one subject), brain damage, right-temporo-occipital area (one subject), postnatal brain damage due to drug reaction (one subject), brain damage due to brain abscess (one subject), prenatal infection, syphilis (one subject) and generalized brain damage (one subject). Subjects in the non-brain damaged group (also twelve males and eight females) had diagnoses of either cultural-familial (eleven subjects) or retardation due to unknown causes (nine subjects). In instances of the last diagnosis, case histories and physical examinations were carefully checked to determine that an organic etiology had, indeed, been rejected.

Each of the forty subjects was administered three successive trials on the Bender Gestalt by the same examiner. Following the procedure of Pacella (1965), a two minute break was given between the second and third trials. The subject was told before the second and third repetitions, "Let's try this again to see how well we do this time." The three trials on 40 subjects yielded a total of 120 separate Bender protocols. These were randomized, numbered 1-120 and Xeroxed so that each of the four scorers would have an identical set of protocols to score.
The scorers were four Masters level psychologists with at least one year post degree testing experience. Each had done extensive diagnostic testing for the special education referrals of public school systems. The Bender Gestalt was a standard test in all of these batteries and was scored by either the Koppitz or Hain system. In short, the scorers had extensive and comparable experience with the Bender Gestalt and the Hain scoring system. Each operated under the assumption that he was scoring the protocols of 120 different subjects. No foreknowledge of the study design was given. They understood that the accuracy of the Hain scoring system was to be evaluated.

Results

Each rater applied Hain's scoring criteria to 120 protocols. To determine the consistency with which the four raters used the Hain's system, rater reliability was estimated via Ebel's method of intraclass correlation (Guilford, 1954, pp. 395 - 397) and the results are presented in Table 1. Ebel's reliability coefficient was .93 which was significant at less than the .001 level. This indicated that these scorers applied the criteria consistently and that variance over the 120 protocols was minimal.
TABLE 1

Ebel's Rater Reliability Summary

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>between person</td>
<td>119</td>
<td>102.12</td>
<td>14.77*</td>
</tr>
<tr>
<td>between rater</td>
<td>3</td>
<td>453.18</td>
<td>65.58*</td>
</tr>
<tr>
<td>remainder</td>
<td>357</td>
<td>6.91</td>
<td></td>
</tr>
</tbody>
</table>

Ebel's reliability .93
*p < .001

While this demonstrates a degree of reliability for the Hain scoring system over a large number of protocols, a note of caution must be made for applying these results to the diagnosis of the individual case. Total score on the Hain system may range from 0 to a maximum of 32. On these 120 protocols, the difference between the lowest and highest score given by the four raters to a particular protocol ranged from 0 to 18, with an average difference of 6.5. Total agreement among raters (a difference of 0) was achieved on only two protocols while, on the other extreme, raters' scores differed by as much as 18 points on one protocol. These reliability difficulties and the attendant implications for the analyses performed in this study are best illustrated by examining the scorers' ratings of several protocols secured with a single subject.
Subject 14 was diagnosed as "brain damage, postnatal infection, pneumonia." Ratings for this subject diverged to the greatest degree among all members of the experimental pool. Reasonable agreement was achieved on trial 1 with respective rater scores of 17, 15, 14 and 14. Scores on trial 2 were 30, 17, 13 and 12, a range of fully 18 points in ratings. Scores on trial 3 were 21, 17, 16 and 10, a range between raters of as much as 11 points. From the Hain studies, the best single cut-off point between brain damaged scores and non-brain damaged scores appeared to be a score of 8. Hain also defined three diagnostic areas with reference to brain damage, a "normal area" of scores from 0 to 5, a "borderline area" of scores from 6 to 12, and a "critical area" of scores of 13 or more. Taking into account the wide range of scores on trials 2 and 3, subject 14 could be considered either "borderline" or very "critical." With the potential of such a wide scoring difference between raters on one protocol, a difference which averaged 6.5 over this sample of 120 protocols, caution must prevail in applying Hain's "areas." In short, while these raters achieved a significant overall reliability with the Hain system over a large N, a diagnostic consensus was somewhat difficult to achieve with the individual case.

The diagnostic validity of the Hain system was evaluated using just the first trial as an approximation
of a standard Bender administration. A "t" test of significance was computed for the differences between brain damaged and non-brain damaged groups. First, the scores of the four raters for each protocol were averaged to obtain one score for each subject. Mean scores for the brain damaged group and the non-brain damaged were 14.12 and 13.42, respectively.

TABLE 2

Means, Standard Deviations and t values for Three Bender Gestalt Trials

<table>
<thead>
<tr>
<th>trial</th>
<th>group</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>brain dam.</td>
<td>20</td>
<td>14.12</td>
<td>4.44</td>
<td>.105*</td>
</tr>
<tr>
<td></td>
<td>non-br. dam.</td>
<td>20</td>
<td>13.42</td>
<td>4.93</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>brain dam.</td>
<td>20</td>
<td>14.38</td>
<td>5.26</td>
<td>.18*</td>
</tr>
<tr>
<td></td>
<td>non-br. dam.</td>
<td>20</td>
<td>12.97</td>
<td>5.73</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>brain dam.</td>
<td>20</td>
<td>14.14</td>
<td>5.18</td>
<td>.24*</td>
</tr>
<tr>
<td></td>
<td>non-br. dam.</td>
<td>20</td>
<td>12.46</td>
<td>4.49</td>
<td></td>
</tr>
</tbody>
</table>

*not significant

Table 2 shows that a nonsignificant "t" value of .105 resulted. Thus, the Hain system was not able to discriminate between clinical groups. For comparison, "t" values were also computed for the remaining two trials and, as Table 2 illustrates, lead to the same conclusion. Any one of the three trials considered alone as an approximation of a
standard Bender administration shows the Hain system unable to discriminate which type of pathology yielded the Bender performance of this sample.

While these results failed to validate the Hain system with this sample, they were consistent in direction with the results predicted and obtained by Pacella (1965) in his similar design. He performed an analysis of variance on his three trials which yielded no significant results on the first two trials as he predicted. Results on his third trial were significant at the .05 level. A comparison of the F per cents for Pacella's three trials with the means and "t" values shown in Table 2 demonstrated a consistent trend with Pacella's sample. F per cents for Pacella's three trials were 1.43, 2.97 and 4.92 (p < .05).

The present study was designed, in part, to evaluate Pacella's prediction of a difference in the non-brain damaged scores with practice over successive trials. He had predicted and received no significant change by the second trial. With this in mind, an analysis of variance within subjects between the first and third trial was performed on this sample in an attempt to more clearly delineate the results. The results of this analysis are summarized in Table 3. The prediction that the brain damaged would not improve due to brain dysfunction was supported. Table 2 shows that brain damaged means were 14.12 on trial one and 14.14 on trial three. The non-brain
damaged were expected to improve by trial three and their trial means do, indeed, demonstrate a slight trend in this direction. Non-brain damaged subjects exhibited means of 13.42 on trial one and 12.46 on trial three. However, the extent of these differences was negligible and the analysis of variance shows no significant difference between trials one and three.

TABLE 3

Analysis of Variance Trials One and Three

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>between subjects</td>
<td>39</td>
<td>11.3</td>
<td>.31*</td>
</tr>
<tr>
<td>subject with groups</td>
<td>38</td>
<td>10.68</td>
<td>.29*</td>
</tr>
<tr>
<td>within subjects</td>
<td>40</td>
<td>34.76</td>
<td>.96*</td>
</tr>
<tr>
<td>residual</td>
<td>38</td>
<td>36.35</td>
<td></td>
</tr>
</tbody>
</table>

*not significant

Discussion

While the Hain scoring system has shown some promise in diagnosing psychiatric patients with brain damage (Hain, 1964; Mosher and Smith, 1965; Kramer and Fenwick, 1966), the findings of this study afford little support for its potential in diagnosing brain damage in adult retardates.
The sample size constitutes one potentially culpable factor. However, Hain used twenty-one brain damaged and twenty-one controls of varying functional diagnoses in his second cross-validation study and Pacella (1965) included twenty-two brain damaged and twenty-two non-brain damaged subjects ranging in age from twelve to thirty-five. Temmer (1965) used twenty-three brain damaged retardates while Smith and Martin (1967) had twenty-five subjects in each group. Thus, the sample size of the present study was consistent with that in previous investigations.

The criteria for the diagnoses may represent an equally critical issue in such Bender Gestalt research with retardates. Feldman's (1953) study addressed itself directly to the question of sample size and the diagnostic criteria involved. He selected his sample from an institution for the retarded with 3400 residents. The experimental pool was initially reduced to 543 residents to secure mature chronological age levels. From these individuals, he could only match 104 familial retardates and 192 non-familial retardates on the basis of age range (within five years) and measured intelligence (five I.Q. points). Any non-familial individuals with specific syndromes such as mongolism, cretinism, phenylpyruvia, etc., were eliminated. For the remaining potential subjects he attempted to reconfirm the original medical diagnoses. Five medical interns administered a special neurological
exam to both groups. Each intern specialized in certain neurological signs and each examined every subject independently. "Hard" neurological signs were distinguished from "soft" or combined signs. The latter was not included unless two or more were present. Feldman had to eliminate many from both groups because only 53% of those originally diagnosed as familials were without signs and as high as 46.5% of those originally diagnosed as non-familials were without neurological signs. Feldman then turned to the case histories and found that the personal and developmental data were "meager and unreliable." Only birth date, order of birth, sex, age at admission and the occupational, academic and institutionalized information on parents and siblings were considered reliable enough to use. Thus, after all of these criteria were met, from the total of 3400 residents he was able to match only 54 endogenous subjects with 54 exogenous subjects. With such a stringent selection process Feldman eliminated over 96% of the potential subject pool. He was confident that only 1.6% were truly retarded for reasons of brain damage.

With this select sample, Feldman used the Pascal and Suttell (1951) scoring system for the Bender Gestalt. The range of scores for both groups were still comparable. Brain damaged subjects ranged from 21 to 271 while non-brain damaged subjects ranged from 26 to 214. Feldman's
statistical comparison did yield significant results at a .05 level of probability but the degree of overlap was still so extensive as to preclude precise applications in individual diagnosis.

Feldman's study reflects many of the difficulties faced in the present study. The forty subjects used were selected from a rather small residential population totaling 531. Based on Feldman's findings only 8.5 subjects (1.6%) would be expected to have supportable brain damage diagnoses. When a specific syndrome associated with retardation was evident as in mongolism or cretenism, the diagnosis was quite simple. However, when the individual exhibited no structural defects, the diagnosis became more complicated.

Difficulties also arose in the available case histories. Family and social histories are important in establishing functional diagnoses (cultural-familial) and organic diagnoses of a genetic origin. Prenatal and postnatal developmental data is an important contributory factor in establishing brain damage diagnoses. In selecting control subjects in the present study for the non-brain damaged group, case histories were certainly not adequate. Only four subjects reportedly had siblings in institutions. This is usually one operational criteria in substantiating cultural-familial retardation. Of the remaining cultural-familial retardates in this group, four
had siblings living at home attending special education classes, the mother of another had three siblings labeled retarded, and two had retarded mothers. Nine with diagnoses of "unknown cause" had to be included to make up the group. Case histories had to be checked very thoroughly to rule out possible organic or brain damage potential. Fully seventeen possible candidates were eliminated by this process. Many of these had little or no medical histories or substantive developmental data to make a determination and, thus, had been labeled retardation due to "unknown cause."

The twenty brain damaged subjects were selected from an initial pool of forty-seven residents so diagnosed. Some of these were immediately eliminated due to low I.Q. or, when tested, could not give a scoruble Bender. Nineteen were eliminated due to diagnostic questions. Each subject had to have had a thorough medical exam with a positive neurological exam before being considered. In addition, some had EEGs or brain scans. Several disease cases were eliminated because developmental medical reports were indirect, sketchy or non-existent. One subject with grand mal seizures and a positive EEG on all tracings was eliminated because of an institutionalized sibling. Of the twenty chosen, seven had examinations by a neurologist, three had EEGs positive on at least 75% of the
tracings, two had brain scans and one had an electrode implant. In effect, only twenty of 531 residents had an adequate medical diagnoses that was positive in relation to brain damage.

This comparison between Feldman's sample size and diagnostic criteria with those of the present study illustrate some critical problems in adult retardate research with the Bender. Feldman found only 1.6% of his subject pool met his diagnostic criteria for brain damage. The present author found only 3.7% brain damaged in his subject pool. Both studies used Bender scoring systems that have demonstrated some reasonable validity with psychiatric patients (Kramer and Fenwick, 1966). Feldman's results with the Pascal and Suttell (1951) system indicated substantial overlap between brain damaged and non-brain damaged. The overlap in the present study with the Hain system is virtually total without any significant discrimination. On the one hand, the criteria for diagnosis in studies of this design may not be adequate. Yet, if they were made more strict with retarded populations, available subjects may be virtually non-existent. On the other hand, when testing adult retardates the problem may lie with the instrument itself or the scoring system used. But before considering the question of the Bender or the system as the culpable factor, one issue remains regarding the diagnosis.
Hain (1964) drew attention to yet another problem, that of general vs. localized brain damage. While his scoring system was able to discriminate brain damage from non-brain damage to some degree, there were several false negatives and a few false positives. 20% of the brain damaged scored five or below and as many as 41% brain damaged scored 8 or below, the best single cut-off point for his sample. Mosher and Smith (1965) found similar results in comparing the Hain system with that of Peek and Quast (1951). As many as 76.8% brain damaged scored 8 or below on the Hain system and were, therefore, falsely classified. Computing a similar cut-off score for the Peek and Quast system, it falsely classified 70.4% of the brain damaged. Incidentally, it should also be noted that both systems falsely classified some of the non-brain damaged (6.7% Hain's system and 5.8% Peek and Quast's system). The problem of distribution overlap was again evident with this sample of psychiatric patients. Yet, comparing both groups statistically, both scoring systems were highly significant at less than the .001 level of chance in their discrimination of brain damaged from non-brain damaged.

Hain (1964) concluded that the Bender with his scoring system could diagnose generally severe impairments of a diffuse nature but could not rule out localized, specific brain damage. Subjects correctly classified in
his studies carried diagnoses such as cerebrovascular insufficiency, arteriosclerosis and diffuse cortical damage. Hain's scoring system failed to pick out subjects with tumors, EEG focal epileptiform activity or those with seizures as the only evidence of brain damage.

In the present study, similar conclusions cannot be drawn with respect to retardates. The two groups were so similar in performance scores that little or no distinction is possible. Moreover, an appropriate cut-off score cannot be computed. If Hain's criteria from his sample were applied to this sample, only two non-brain damaged subjects would score 8 or below, leaving 90% of this non-brain damaged group falsely classified. Thus, the criteria appear valueless when extrapolated to an adult sample of mentally retarded subjects. Moreover, as has been seen, Hain's system does not even distinguish retarded subjects with generalized brain damage from those presumedly devoid of such pathology.

Hain considered the problem to lie more with the Bender Gestalt than with his scoring system.

False negatives are to be expected from any single test that measures only one or a few dimensions of impairment associated with brain damage. Impairment associated with brain damage is dependent upon many factors, including the brain lesion's location, duration, extent, and etiology. No single test can be expected to tap impairment associated with all types of brain damage (Hain, 1964, p. 40).

Mosher and Smith (1965) agreed that the problem lies with
the instrument. They make an even stronger case while disagreeing with Hain in part.

Contrary to Hain's suggestion that false negatives on the test are composed of patients exhibiting small localized brain lesions, inspection of the test misses in this sample indicated that false negatives were not confined to this group. Patients with subdural hematomas, cerebral vascular accidents, arteriosclerosis, and diffuse degenerative brain disorders, which were established by surgical reports or special radiological techniques such as pneumoencephalography, were also test misses.... Moreover, the cases of traumatic encephalopathy which were classified correctly tended to be the severely injured cases requiring surgery or having skull fractures rather than the concussion cases. It appears that the brain damaged patients which are classified correctly as true positives on the BGT are most often cases of obvious brain injury which do not require psychological assistance to diagnose (Mosher and Smith, 1965, p. 535).

If the imprecision of the Bender Gestalt per se is the culpable factor in the generally poor diagnosis of brain damage, the findings of the present study and those preceding it at least imply even greater difficulty with adult mental retardates than other populations. The results of a one trial administration in the present investigation and those of Feldman (1953) and Pacella (1965) showed little or no significance. By comparison, greater significance is found with a standard administration using adult psychiatric patients (Griffith and Taylor, 1960; Hain, 1964; Niebuhr and Cohen, 1956; Reznikoff and Olin, 1957; Tolor, 1956, 1958). There appears to be a greater degree of distribution overlap in the studies using adult retardates. Furthermore,
Canter's BIP strongly suggested a means for overcoming this problem with psychiatric patients, but failed to do so when applied to retarded children and adults (Canter, 1966; Adams, 1968, Song and Song, 1969). In effect, the cautionary note of several authors (Billingslea, 1963; Hain, 1964; Mosher and Smith, 1965; Niebuhr and Cohen, 1956) that care must be exercised when diagnosing the individual case, and that the Bender Gestalt cannot rule out brain damage, may be especially true in cases of retardation.

If, indeed, the problem lies with the Bender itself, it seems safe to conclude that applying any scoring system to the protocols may confound the issue. The specific problem may be best described as one of diagnostic reductionism. For example, when a neurologist diagnoses, he first conducts a preliminary neurological exam. On the basis of this he may suspect or diagnose brain damage. If he only suspects it, he may choose to employ more sophisticated tools such as an EEG, brain scan or even an arteriogram. On the other hand, if his preliminary exam indicates no signs of brain damage he may rule this diagnosis out entirely. But it is possible that he could very well miss something another instrument would pick up. The full range of potential detectors are only applied if more gross indices include some positive signs. The problem with a preliminary neurological
exam is analogous to that encountered with the Bender Gestalt. The Bender Gestalt can diagnose some brain damage but is not adequate for ruling it out (Billingslea, 1963; Hain, 1964; Mosher and Smith, 1965). Like the standard neurological exam, the Bender Gestalt yields limited information for diagnosis. When delimited to a single scoring system, the amount of diagnostic information is reduced to one raw or z score. While this standardizes conclusions to be drawn, it also has the effect of reducing the sensitivity of the instrument as an initial screening device for further, more precise evaluative measures.

The present study also attempted to cross-validate the results of Pacella's (1965) study. Pacella corroborated predictions that subjects devoid of brain damage would exhibit practice-mediated improvements with successive administrations of the Bender while brain damaged subjects would not benefit from such practice. The present study revealed a similar but nonsignificant trend as Table 2 illustrates. Whether the failure to replicate Pacella's findings reflects imprecision in the hypothesis or the methodology cannot be determined from the present study. However, this investigator's speculations would emphasize the latter. It may well be that such effects would be more probable if additional trials were incorporated to magnify any potential gains
and some form of trial-by-trial feedback and/or reinforcement were incorporated to maximize optimal conditions for learning to occur. This author would speculate further that such a design may, in fact, improve results over those in this or Pacella's study. In any event, when the preceding possible culpable factors are considered, namely, limited sample size, overly gross diagnostic validity criteria, the imprecision of the Bender test itself and the apparent insensitivity of the Hain scoring system, it would seem illogical to conclude that the Pacella hypothesis has been definitively rejected.
CHAPTER IV

SUMMARY AND CONCLUSION

In the early 1930's, Lauretta Bender began developing a test of visual-motor gestalt functioning based on the designs of Wertheimer (1923). Since the publication of her first monograph (Bender, 1938), the Bender Gestalt has blossomed into one of the most widely used psychological tools for clinical diagnosis (Sundberg, 1961). The research literature has demonstrated mixed results with respect to the Bender's diagnostic latitude and proficiency. This simple instrument consisting of nine geometric designs for freehand copying has been used as a diagnostic tool with hospitalized mental patients having a variety of pathologies, with school age children as a test of visual-motor proficiency, intelligence and emotional problems, with retardates as a test of organic brain pathology and with all types of patients as a projective device.

Attempts to refine the Bender Gestalt have generally taken two directions: 1) the development of standardized scoring systems and 2) modifications of the administrative procedure. Bender herself used a clinical inspection method
to evaluate protocols. The subjective judgments of the experienced clinician have continued as an acceptable means of interpretation (Billingslea, 1963). Bender (1938, 1946) also supplied age appropriate standards by which others could judge the proficiency of subjects' reproductions. This crude scoring system inspired the development of several more sophisticated methods aimed at objectivity and standardized conclusions. Billingslea (1948) published the first of these, but it was considered too cumbersome for practical use. Peek and Quast (1951) engineered a well designed system but its impact has been minimal due to its lack of availability (Billingslea, 1963). The most influential scoring system has been that of Pascal and Suttell (1951). This system has been the most widely used in research and established a pattern for devising other scoring methods. Koppitz (1958, 1960, 1962, 1964) modified Pascal and Suttell's system to be used with children. Koppitz's "Developmental Bender Scoring System" is probably the most widely used system in clinical practice. Hain (1964) constructed a system for accentuating the Bender's diagnostic proficiency with brain damage. The Hutt Adaptation of the Bender Gestalt Test (1969) encompassed both types of refinement. In fact, Hutt modified the Bender in three ways. He developed his own set of stimulus cards, a three part administrative
procedure and a scoring system. The Hutt method was intended to serve primarily projective interpretation. Methods modifying the administrative procedure were directed especially toward refining the Bender's ability to diagnose brain damage. Besides Hutt (1969), these administrative modifications include the recall protocol developed by Sullivan and Welch (1948), the Background Interference Procedure (BIP) designed by Canter (1966) and the three trial administration proposed by Pacella (1965).

The present investigation was designed to evaluate two of these refinements, one in each category, the Hain (1964) scoring system and the Pacella (1965) three trial administration procedure. Both refinements were designed to magnify Bender discrimination of brain damage. The Hain scoring system was developed with and validated on brain damaged psychiatric patients. The first purpose of the present study was to evaluate the reliability and validity of this system with adult mental retardates. Pacella's (1965) findings suggested that his three trial administration would demonstrate performance improvement by the third trial for non-brain damaged retardates due to practice and learning. Brain damaged subjects would not show improvement due to the nature of their impairment. The second purpose of the present investigation was to cross-validate the results
of Pacella's study.

Twenty brain damaged and twenty non-brain damaged adult mental retardates matched in age and I.Q., were administered three successive trials of the Bender Gestalt by the same examiner. Following Pacella's direction, a two minute break was interspersed between the second and third trial. The resulting 120 protocols were randomized and scored independently by four experienced raters using the Hain scoring system. Ebel's rater reliability formula was applied to the four raters' scores and showed the Hain system could be consistently applied by different raters to the protocols of adult retardates. Closer examination of Hain raw scores for individual protocols, however, revealed an average difference between highest and lowest score of 6.5. This difference demonstrated that difficulty may be encountered in applying Hain's diagnostic criteria to the individual case. The first trial was taken as an approximation of a standard administration and differences between brain damaged and non-brain damaged subjects were analysed by computing a "t" score. A non-significant "t" value of .105 failed to confirm the diagnostic validity of the Hain scoring system with this sample. The difference in trial means and "t" values between the first and third trials exhibited results in the same direction as those of Pacella, but an analysis of variance showed this difference to be nonsignificant.
Five factors were considered as probable determinants of the nonsignificant findings of this investigation: 1) sample size, 2) diagnostic criteria used in subject selection, 3) the Bender Gestalt test itself, 4) the scoring system used, and 5) the minimized practice benefits produced in testing the Pacella procedure. Sample size was found to be consistent with previous studies using both psychiatric patients or retardates. Diagnostic criteria were also considered to be consistent with previous investigations and strict enough in that only 3.7% of the potential subjects were selected for the brain damaged group. The inappropriateness of Bender Gestalt itself for the mentally retarded was considered to be the most culpable factor. When results with psychiatric patients were compared with results with retarded populations including the present study, only one well designed investigation (Feldman, 1953) has achieved significant results with adult mental retardates on a one trial administration. Neither Pacella (1965) nor the present author found significant differences between groups with a one trial administration. It was concluded that, if the failure of the Bender is, in fact, the contributing variable, then any scoring system used would not improve prediction. Instead, such objective scoring systems would actually reduce the data available for the final diagnostic judgment. Findings of the present
investigation were consistent in direction but nonsignificant in support of Pacella's (1965) hypothesis. While no substantive conclusions could be drawn, the present author speculated that the lack of significant results was due to limitations in Pacella's method and those of the present study. It was suggested that additional trials combined with feedback and/or reinforcement might maximize learning conditions and magnify potential gains.

In the final analysis, the findings of this investigation lead this author to confirm the conclusions of many of his predecessors regarding the Bender Gestalt. First, the Bender Gestalt may serve to some extent as a screening device for detecting brain damage but cannot be used to rule it out. Furthermore, it cannot be used in and of itself to conclude that brain damage does, in fact, exist. Moreover, an examiner must exercise even more caution when addressing brain damage in a mentally retarded adult than in other populations. This author would, therefore, support wholeheartedly the following conclusion of Mosher and Smith.

It is not sufficient to indicate that the BGT scores are significantly different for a brain damaged and a control group. It is necessary to demonstrate satisfactory prediction in individual cases and to examine the incremental validity of the BGT scoring systems. Does the BGT add some increment to the reduction of serious errors which are frequently made? A comparison
of the clinician's subjective confidence in the BGT in distinguishing brain damaged from nonorganic patients suggests that the clinician's confidence is misplaced and may be a source of serious error (Mosher and Smith, 1965, p. 334).
REFERENCES


Baroff, G. Bender Gestalt visuo-motor function in mental defectives. *American Journal of Mental Deficiency*, 1957, 61, 753-760.


Canter, A. A background interference procedure to increase sensitivity to the Bender Gestalt Test to organic disorder. Journal of Consulting Psychology, 1966, 30, 91-97.


Temmer, H. W. Wechsler intelligence scores and Bender Gestalt performance in adult male mental defectives. *American Journal of Mental Deficiency*, 1965, 70, 142-147.


APPENDIX

Hain Scoring System Raw Scores of Four Raters
On Three Trials of the
Bender Gestalt
# Brain Damage Group

<table>
<thead>
<tr>
<th>Subject Number</th>
<th>Trial 1 Rater Scores</th>
<th>Trial 2 Rater Scores</th>
<th>Trial 3 Rater Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R₁ R₂ R₃ R₄</td>
<td>R₁ R₂ R₃ R₄</td>
<td>R₁ R₂ R₃ R₄</td>
</tr>
<tr>
<td>1</td>
<td>11 13 5 11</td>
<td>18 13 12 14</td>
<td>16 9 9 11</td>
</tr>
<tr>
<td>2</td>
<td>14 7 9 7</td>
<td>8 5 9 6</td>
<td>9 6 6 6</td>
</tr>
<tr>
<td>3</td>
<td>19 23 19 18</td>
<td>24 18 25 24</td>
<td>29 24 26 26</td>
</tr>
<tr>
<td>4</td>
<td>20 14 18 18</td>
<td>23 14 18 18</td>
<td>23 14 15 15</td>
</tr>
<tr>
<td>5</td>
<td>15 6 6 9</td>
<td>12 15 9 10</td>
<td>10 7 9 6</td>
</tr>
<tr>
<td>6</td>
<td>18 13 12 13</td>
<td>17 10 11 12</td>
<td>14 10 10 14</td>
</tr>
<tr>
<td>7</td>
<td>20 13 22 13</td>
<td>19 18 13 18</td>
<td>19 16 18 18</td>
</tr>
<tr>
<td>8</td>
<td>26 14 18 17</td>
<td>25 18 20 21</td>
<td>21 14 21 17</td>
</tr>
<tr>
<td>9</td>
<td>32 15 20 16</td>
<td>24 20 16 16</td>
<td>27 16 15 15</td>
</tr>
<tr>
<td>10</td>
<td>14 9 9 9</td>
<td>9 14 10 10</td>
<td>7 5 10 5</td>
</tr>
<tr>
<td>11</td>
<td>15 8 9 8</td>
<td>14 13 8 11</td>
<td>16 10 7 10</td>
</tr>
<tr>
<td>12</td>
<td>17 13 17 13</td>
<td>23 17 18 18</td>
<td>26 15 12 16</td>
</tr>
<tr>
<td>13</td>
<td>17 14 15 14</td>
<td>30 12 17 13</td>
<td>21 17 10 16</td>
</tr>
<tr>
<td>14</td>
<td>19 14 13 19</td>
<td>16 15 17 16</td>
<td>15 15 18 18</td>
</tr>
<tr>
<td>15</td>
<td>21 20 23 25</td>
<td>23 14 22 28</td>
<td>18 13 18 10</td>
</tr>
<tr>
<td>16</td>
<td>19 15 17 18</td>
<td>23 14 18 19</td>
<td>21 20 18 20</td>
</tr>
<tr>
<td>17</td>
<td>8 7 9 8</td>
<td>14 6 10 10</td>
<td>14 21 18 15</td>
</tr>
<tr>
<td>18</td>
<td>15 14 16 12</td>
<td>10 9 7 9</td>
<td>15 8 11 10</td>
</tr>
<tr>
<td>19</td>
<td>9 10 11 10</td>
<td>7 7 7 7</td>
<td>10 13 10 9</td>
</tr>
<tr>
<td>20</td>
<td>9 10 6 6</td>
<td>7 6 4 6</td>
<td>8 9 5 7</td>
</tr>
</tbody>
</table>
## Non-Brain Damage Group

<table>
<thead>
<tr>
<th>Subject Number</th>
<th>Rater Scores</th>
<th>Rater Scores</th>
<th>Rater Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R&lt;sub&gt;1&lt;/sub&gt;</td>
<td>R&lt;sub&gt;2&lt;/sub&gt;</td>
<td>R&lt;sub&gt;3&lt;/sub&gt;</td>
</tr>
<tr>
<td>21</td>
<td>23 17 23 23</td>
<td>18 10 15 11</td>
<td>24 10 19 17</td>
</tr>
<tr>
<td>22</td>
<td>13 13 13 13</td>
<td>19 13 14 15</td>
<td>14 10 11 13</td>
</tr>
<tr>
<td>23</td>
<td>13 11 10 10</td>
<td>10 9 4 10</td>
<td>12 10 7 9</td>
</tr>
<tr>
<td>24</td>
<td>10 3 4 3</td>
<td>10 4 0 0</td>
<td>14 8 9 9</td>
</tr>
<tr>
<td>25</td>
<td>12 13 6 13</td>
<td>10 18 10 10</td>
<td>6 10 9 6</td>
</tr>
<tr>
<td>26</td>
<td>13 6 2 2</td>
<td>8 4 0 3</td>
<td>8 5 2 4</td>
</tr>
<tr>
<td>27</td>
<td>22 13 21 19</td>
<td>28 21 25 21</td>
<td>23 19 20 20</td>
</tr>
<tr>
<td>28</td>
<td>27 15 15 17</td>
<td>28 11 21 17</td>
<td>13 14 10 10</td>
</tr>
<tr>
<td>29</td>
<td>27 17 11 21</td>
<td>26 15 15 15</td>
<td>22 13 12 17</td>
</tr>
<tr>
<td>30</td>
<td>20 16 11 14</td>
<td>18 16 12 17</td>
<td>19 11 9 10</td>
</tr>
<tr>
<td>31</td>
<td>21 10 15 11</td>
<td>15 16 14 15</td>
<td>19 14 19 19</td>
</tr>
<tr>
<td>32</td>
<td>14 9 14 14</td>
<td>15 13 16 15</td>
<td>17 16 18 15</td>
</tr>
<tr>
<td>33</td>
<td>10 6 7 6</td>
<td>8 4 7 7</td>
<td>8 7 7 7</td>
</tr>
<tr>
<td>34</td>
<td>17 16 16 17</td>
<td>16 12 18 18</td>
<td>15 14 14 14</td>
</tr>
<tr>
<td>35</td>
<td>17 11 19 15</td>
<td>20 15 19 17</td>
<td>20 13 22 13</td>
</tr>
<tr>
<td>36</td>
<td>19 17 10 13</td>
<td>17 7 10 14</td>
<td>16 12 7 16</td>
</tr>
<tr>
<td>37</td>
<td>19 11 15 13</td>
<td>12 9 10 12</td>
<td>13 7 9 10</td>
</tr>
<tr>
<td>38</td>
<td>17 13 24 16</td>
<td>20 16 24 20</td>
<td>16 18 17 17</td>
</tr>
<tr>
<td>39</td>
<td>13 17 14 13</td>
<td>17 9 12 12</td>
<td>14 10 10 12</td>
</tr>
<tr>
<td>40</td>
<td>3 2 3 2</td>
<td>2 2 0 2</td>
<td>3 7 1 3</td>
</tr>
</tbody>
</table>