

https://dspace.org/

Inter and intra-rater reliability of the scoring and analysis of Bender Gestalt and Human Figure Drawing protocols in a clinical setting

Suttner, Vivien

2014-04-03
http://hdl.handle.net10539/14480

Downloaded from a DSpace repository. https://dspace.org/

INTER AND INTRA-RATER RELIABILITY OF THE SCORING AND ANALYSIS OF BENDER GESTALT AND HUMAN FIGURE DRAWING PROTOCOLS IN A CLINICAL SETTING

VIVIEN SUTTNER

A research report submitted to the Faculty of Specialised Education, University of the Witwatersrand, in partial fulfilment of the requirements for the Degree of Master of Education (Educational Psychology)

Johannesburg, 2000.

Degree awarded with distinction on 27 June 2000

ABSTRACT

The Bender Gestalt and the analysis of the Human Figure Drawing are amongst the most widely used assessment tools in South Africa. This research examines the inter and intra-rater reliability of scores and diagnoses made by two clinicians from a clinical sample of 104 Bender Gestalt and Human Figure Drawing protocols. Two scoring sessions were held, with a three week interval between them. The protocols were scored according to the Koppitz criteria. On the basis of these results the clinicians placed each child into one of four proscribed diagnostic categories: Exhibiting visual-perceptual difficulties; emotional difficulties; both difficulties; neither difficulties. The results indicated significant inter-rater differences between the raters' diagnoses, and scoring of the Bender Gestalt. With regards to intra-rater differences, i.e. individual differences from session one to two, significant differences were evident in the scoring of the Bender Gestalt. No significant intra-rater differences emerged on the scoring of the Human Figure Drawing. These results indicate that there is cause for concern regarding the reliability of clinicians' interpretations based on The Bender Gestalt and Human Figure Drawings.

KEY WORDS

Bender Gestalt, Human Figure Drawing (HFD), Inter- and Intra-Rater Reliability, Clinicians, Diagnosis, Visual-Perceptual Assessment, Emotional Assessment, Projective Testing

DECLARATION

I hereby declare that this thesis is my own, unaided work. It is being submitted for the degree of *Master of Education* (Educational Psychology) at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.



Vivien Suttner

DEDICATION

This research is dedicated to the memory of my father, Geoffrey Leveson, who valued academic learning and intellectual pursuits.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the following people for their contributions towards this research.

- My supervisor, Lynne Cornfield, for her guidance, valuable insight, patience and understanding throughout the process of the creation of this research report.
- My husband, Immanuel Suttner, for his unconditional and tireless love and support, valuable and insightful guidance, astute grasp of the subject matter and accurate proof-reading.
- My mother, Marcia Leveson, for her tireless encouragement and support and valuable assistance in typing and proof-reading.
- John Mansfield, for his advice, support and willingness with the statistical analyses of the research.
- Max Weldon and Rosanne Green, for undertaking to participate in the study.
- The Bryanwood Therapy and Assessment Centre for allowing me to use data for the research.
- Lesley Caplan for her assistance in the formulation of the research aims and her insight and guidance over many years.
- The Life Training Programme for the contribution it has made to enriching my life.
- The financial assistance of the Centre for Science Development towards this research is hereby acknowledged.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION 1

CHAPTER TWO: LITERATURE REVIEW 3

2.1 Psychological testing and psychological measurement 3

2.2 The role of psychological assessment in schools 5

2.3 Assessment of visual-perceptual skills 6

2.4 Assessment of emotional functioning based on the human figure drawing 13

CHAPTER THREE: RATIONALE AND AIMS OF THE STUDY 13

3.1 Rationale for the study 19

3.2 Aims of the study 24

CHAPTER FOUR: METHOD 25

4.1 Subjects 25

4.2 Sample 25

4.3 Sampling procedure 25

4.4 Age at time of test 25

4.5 Distribution of rater by test 26

4.6 Instruments 27

4.6.1 Bender Gestalt Test of Visual Motor Integration 27

4.6.2 The Human Figure Drawing 27

4.7 Procedures 28

4.8 Specific Hypotheses 30

4.9 Statistical Procedures 30

CHAPTER FIVE: RESULTS 34

5.1 Inter-rater reliability of the scoring of the Bender Gestalt 34

5.2 Intra-rater reliability of the scoring of the Bender Gestalt 37

5.3 Inter and intra-rater reliability of the scoring of the HFD 39

5.4 Intra-rater reliability of the scoring of the HFD 41

5.5 Statistics for intra-rater agreement of the diagnoses of the child's difficulty 42

5.6 Inter-rater agreement of the diagnosis of the child's difficulty 46

CHAPTER SIX: DISCUSSION 52

6.1 Introduction 52

6.2 Inter-rater reliability in the use of the Bender Gestalt Test 53

6.3 Intra-rater reliability in the use of the Bender Gestalt Test 55

6.4 Inter-rater agreement - in the use of the Human Figure Drawing 56

6.5 Intra-rater reliability of the diagnosis of the child's difficulty 57

6.6 Inter-rater agreement in diagnosis 61

6.7 Limitations of the study and suggestions for future possible research 64

6.8 Conclusion 68

REFERENCE LIST 69

APPENDICES

LIST OF FIGURES

Figure 1.	Age distribution	26
Figure 2.	Bender correlations across rater by test	38
Figure 3.	HFD correlations across test by rater	41
Figure 4.	Pie chart: Test 1 vs. Test 2	43
Figure 5.	Pie chart: Rater M vs. Rater R	48

LIST OF TABLES

Table 1.	Summary statistics for age at test	26
Table 2.	Distribution of rater by test	26
Table 3.	Summary statistics – Bender Gestalt Test - inter-rater comparison	34
Table 4.	Analysis of variance for dependent variable BMA	35
Table 5.	Analysis of variance for factors modelling BMA	35
Table 6.	Analysis of Variance - Comparison between BMA means per rater	35
Table 7.	Pearson's Correlation Coefficient – Bender Gestalt inter-ratercomparison	36
Table 8.	Summary statistics – Bender intra-rater comparison	37
Table 9.	Pearson's Correlation Coefficient – Bender intra-rater comparison	38
Table 10.	Analysis of variance for dependent variable HFD	39
Table 11.	Correlation analysis – HFD inter-rater comparison	40
Table 12.	Correlation analysis – HFD intra-rater comparison	41

Table 13. Contingency table - comparison of analysis by test session 42

Table 14. Statistics for intra-rater agreement - Chi Square, Likelihood Ratio Chi-Square, Fisher's Exact Test, Cohen's Kappa 46

Table 15. Contingency table - comparison of analysis by rater 47

Table 16. Statistics for inter-rater agreement Chi Square, Likelihood Ratio Chi Square, Fisher's Exact Test, Cohen's Kappa 51

CHAPTER ONE: INTRODUCTION

For more than a hundred years psychologists and researchers have attempted to measure human personality characteristics. Over time researchers grounded in different theoretical bases have proposed various assessment models. The practical manifestations of these models have undergone numerous incarnations. However, the essential aims of enabling the researcher and clinician to understand, control and predict behaviour appear to underlie all approaches to psychological measurement (Reichenberg & Rafael 1992). The course of psychological measurement has, however, been fraught with controversy concerning its philosophical soundness and the practical usefulness of the techniques that emerge from the various approaches. In particular, the validity and reliability of assessment tools have been called into question. Tools emerge from theoretical approaches to the understanding of mind and behaviour. Once the tool is "on the market", researchers and clinicians "test" these tools and a proliferation of data results. This data generally contains various conclusions as to the ongoing utility of tools, and may propose variations or improvements.

This research focuses on two of the most frequently used assessment tools in clinical practice: The Bender Gestalt Test of Visual Motor Integration and the analysis of the Human Figure Drawing (hereafter referred to as the Bender Gestalt and the HFD). Research into these tools is particularly significant in the South African context, where psychological assessment often occurs in a resource poor environment, and a child's difficulty and referral may be decided using only a handful of tools, the Bender Gestalt and HFD being prominent amongst these. Inaccurate identification of a child's difficulty could have serious ongoing repercussions for him/her. It is therefore important to establish that clinical formulations based upon these tools are sufficiently consistent.

Although the Bender Gestalt and HFD tools have been frequently maligned in the literature as ineffective and subjective, little research has been conducted into the reliability of the clinical formulations derived from them. This research argues that the effectiveness of these tools lies in the hands of the clinician. It therefore examines the inter and intra-rater reliability of these tools when used to assess primary school children.

Two research psychologists, each with several years of clinical experience, were selected as the subjects for the research. They were asked to score 52 Bender Gestalt and 52 HFD protocols that had been gathered from a clinical sample. Thereafter they were asked to diagnose the child into one of four specific diagnostic categories: Exhibiting primarily visual-perceptual difficulties; primarily emotional difficulties; both difficulties; neither difficulties. The data thus obtained were subjected to a number of statistical analyses. These are presented, along with a discussion of the results, in subsequent chapters. Limitations of the study and suggestions for future research conclude this report.

CHAPTER TWO: LITERATURE REVIEW

2.1 Psychological testing and psychological assessment

The practice of assessment as a systematic activity is fairly recent and has its primary historical antecedent in the theory and practice of medicine. The central pre-occupation of the latter discipline has been the exact description of symptoms of illness and detailed recordings of treatment processes. Emerging from the medical paradigm, psychological measurement, since its scientific beginnings in the 19th century, has been centrally concerned with the description and measurement of individual differences (Anastasi, 1982), as well as the development and evaluation of methods aiming to apply relatively stable generalisations to human behaviour (ibid, 1982). Over time, psychological measurement has been formalised and expanded into a process more generally referred to by psychologists as "assessment" and is regarded as a highly complex field.

Psychological testing and psychological assessment are often considered to be synonymous. However, the terms refer respectively to broader and narrower functions. Psychological testing indicates the technical process of administering and scoring a "psychologically relevant instrument" (Gabel, Gerald & Butnik, 1986, p.3). Psychological assessment indicates the process of integrating and understanding the results of all components of the formal testing, the individual's circumstantial and developmental history, observations of the individual's behaviour during the assessment, as well as interviews with relevant parties, e.g. parents and teachers (ibid). The goal is to quantify, interpret and understand the nature of the child's difficulty in order to assist him or her in the relevant area(s).

Over the past 80 years, the diverse disciplines within the field of psychology have spawned a plethora of tools that endeavour to quantify and explicate various aspects of human psychological functioning. Neuropsychologists have been interested in aspects such as brain damage and changes to the brain over time and have developed various tests of cognition, perception and motor functioning.

The psychoanalytic movement has developed a number of tools attempting to tap unconscious processes. The intelligence testing movement (the pioneering work behind the modern testing movement) resulted in the development of tools aimed at isolating various types of intelligence (Cox, 1993). Today there are hundreds of standardised tests claiming to provide objective measures for use by the clinician in his or her attempt to clearly articulate an individual's level of functioning in various spheres. Clinicians depend on these tools to provide appropriate normative data, and frequently use them to form the basis for either a formal diagnosis of the individual's difficulty, or the basis for a particular intervention. Along with these tools is an equally substantial body of research examining the theoretical assumptions underlying these tools as well as their practical validity and reliability.

While formal psychometric tools or tests are samples of current functioning, the clinician uses these tools to extrapolate to an estimate of typical functioning (Chittooran & Miller, 1998). With this generalisation being made, it is clear that the ongoing research into the aims, processes and results of assessment tools is of the utmost importance. It is also vital that clinicians are abreast with the ongoing research into the validity and reliability of tools and are aware of the implications of making diagnoses on the basis of the results of (in some cases possibly questionable) assessment tools.

There are an abundance of assessment tools in the psychologist's armamentarium, with researchers claiming that certain devices tap a particular aspect of psychological functioning. The Bender Gestalt and Human Figure Drawing or (Draw a Person Test) are two of the most frequently used assessment tools in clinical practice. A study by Wilson and Reschley (1996) established that the Bender Gestalt ranked second to the Wechsler Scales in terms of usage by school psychologists, while the HFD ranked third. These tools have been in use since the 1920's and, despite extensive and ongoing debates as to their psychometric soundness and clinical utility, are still heavily relied upon and almost always included in an assessment battery.

The Bender Gestalt and HFD have their origins in different branches of psychology. The Bender Gestalt emerged from neurospsychology, while the HFD is more commonly associated with the psychoanalytic movement. These two short, pencil and paper tests are almost always included in a standard battery to gain insight into the child with a wide range of presenting difficulties.

2.2 The role of psychological assessment in schools

The accurate assessment of children's functioning is a critical concern to teachers and educational psychologists (Chittooran & Miller, 1998). Indeed, in the United States and the United Kingdom schools are among the largest test users, with clinicians devoting a significant amount of time to the administration and interpretation of assessment tools (Anastasi 1982). Surveys of school psychologists in the United States indicate that assessment accounts for at least 50% of their professional practice (Chittooran & Miller, 1998). No formal statistics appear to be available for the South African situation but, based on the researcher's experience in both private and governmental contexts, it seems reasonable to assume that a primary school child experiencing learning difficulties has a high possibility of undergoing a psychological assessment of some sort in his/her school career.

It is apparent that clinicians place considerable emphasis on assessment tools as part of a general problem-solving process in the explication of children's difficulties. While these tools are regarded as only one method of obtaining data, they are often viewed as the most fundamental component of an assessment process (as compared to interviews with parents and teachers or observations of the child's behaviour). This is because they are a time efficient, focused and hopefully objective means of obtaining data (Gabel et al, 1986). Indeed, where the time and budget for the clarification of a child's difficulties is significantly limited, as for example in contemporary South Africa, a reliance on test data increases dramatically.

As mentioned, because diagnoses and interventions which directly affect a child's future are significantly informed by and based upon the results of these tests, it is of the utmost importance that the tools demonstrate high degrees of reliability and validity. This has become even more imperative in recent years because of the vocal body of researchers calling for testing to be linked more closely to remediation, as part of a general process of bringing theory and practice closer together. It is therefore vital that tests actually measure what they purport to measure, i.e. have a high degree of construct validity, and that they effectively are able to predict future performance, i.e. demonstrate high predictive reliability. An essential aspect of their reliability is that there is minimal variation in results obtained for the same child's test protocols by different clinicians, i.e. for significant inter-rater reliability to be evident. Given the relative lack of recent research in the latter area, and the possible suitability of the Bender Gestalt and HFD in the South African context, this report has chosen to explore the inter-rater reliability of these tools as one facet of their general reliability.

Educational psychologists[1] in South Africa at present appear to use The Bender Gestalt primarily to assess visual-perceptual-motor functioning and the HFD to gain insight into emotional functioning and behavioural adjustment. A background to the assessment of visual-perceptual-motor functioning and the use of the human figure drawing as a measure of emotional functioning will now be more fully explored.

2.3 Assessment Of Visual-Perceptual Skills

Adequate perceptual-motor development is considered by many educators and psychologists to be a prerequisite for the development of academic skills [2] (Salvia & Ysseldyke, 1988). Children who perform poorly on visual-perceptual tasks are said to demonstrate difficulties in the visual-perceptual domain and these difficulties are thought to contribute to, or even cause, learning problems (Fuller, Awadh & Vance, 1998).

---

[1] This assumption is based on the researcher's experience in various South African settings

[2] Perceptual-motor development is defined here as the act of interpreting a stimulus registered in the brain by the visual sense and responding to the stimulus through a physical (motor) activity (Drever, 1979).

The adequate acquisition of visual-perceptual skills is considered by many authors to relate in particular to the effective acquisition of reading, writing and arithmetic skills (Koppitz, 1968).

The Bender Gestalt - which was originally developed to diagnose brain injury - has, since the late 1930's, become fashionable as a tool to assess children experiencing academic difficulties, in order to establish the extent to which perceptual-motor difficulties may underlie their academic difficulties (Salvia & Ysseldyke, 1988). The apparent relationship between perceptual-motor functioning and school achievement means that devices tapping the perceptual-motor area are considered by clinicians to be highly relevant to the assessment of children in terms of both problem identification and the development of remediation techniques.

2.3.1 Theoretical foundations of visual-perceptual assessment

The assessment of perceptual-visual-motor skills is deeply rooted in the theoretical and conceptual literature of Gestalt Psychology (Tolor & Schulberg, 1962). This approach focuses on the perceptual reaction of the human organism and proposes four principles of perception (Fuller et al, 1998):

(1) *Inhomogeneity,* or the relationship of figure and ground to perception - in order for a figure to be seen, the ground must be *inhomogenous* to the figure (a black figure cannot be seen against a black background).
(2) *Interaction of figure and ground* - Koffka (1935) demonstrated that variations of the ground influence one's perception of the figure.
(3) *The laws of grouping or closure* - these imply that each figure has its own properties and when certain conditions (proximity, similarity, common fate) are obtained between the parts, a whole cohesive figure is perceived. The greater the degree of the conditions, the more stable the perceived object is to be.
(4) *Wertheimer's Law of Pragnanz* - this principle states that there is a tendency to get a decisive *structuralisation* of the perceived object with true orientation. In perceptual-motor assessment the whole is greater than the sum of the parts, so a perceived stimulus (diamond, square, triangle or letter of the alphabet) has greater meaning for the individual than its component parts (Bender 1938,Fuller et al 1998).

2.3.2 *The Gestalt function*

The Gestalt function is defined by Bender (1938, p. 35) as, "the phenomenon whereby the human organism responds to a given constellation of stimuli as a whole; this response is referred to as a pattern, or gestalt." According to Bender (1938) all integrative processes within the nervous system occur in these constellations.

The founding tenet of Gestalt Psychology is that organised units or structured configurations are the primary forms of biological reaction at the psychological level of arousal and that, in the sensory field, these organised units correspond to configurations in the environment (Bender, 1938).

Hallahan and Cruickshank, 1973 (cited in Salvia & Ysseldyke, 1988) traced the history of the study of visual-perceptual problems in mentally challenged, brain-injured and learning disabled children. These authors followed the historical roots of current practices in visual-perceptual assessment to the early work of Goldstein, Werner and Strauss. Goldstein had been studying World War 1 soldiers who had suffered traumatic head injuries. He observed that these patients exhibited the psychological characteristics of perseveration, figure background confusion and concrete behaviour.

In the mid 1940's the German psychologists Werner and Strauss began to study the behavioural pathology evidenced by brain-injured individuals. They wanted to establish whether the psychological manifestations of brain injury found in adults by Goldstein was also observable in children. This was an important shift from the extensive history of interest in perception and perceptual problems in adults and the brain-injured, to a particular concern for the perceptual and motor problems of non brain-injured children who did not perform in the academic arena (Salvia & Ysseldyke, 1988).

Therefore, while visual-perceptual devices have been used for some time to diagnose brain injury, from the mid 1940's there was a significant development in the use of various visual-perceptual devices to diagnose learning disabilities. The assumption underlying this shift is that the child experiencing visual-perceptual disability is likely to find academic learning difficult and, as a "by product", is likely to fail to generally adjust to the social and emotional demands of the classroom (Fuller et al 1998, Salvia & Ysseldyke 1988). Research has found that these difficulties, regardless of etiology, can be ameliorated by specific training (Salvia & Ysseldyke, 1988). It is therefore believed that the accurate identification and training of children with visual-perceptual difficulties can help prevent many instances of school failure and maladjustment.

2.3.4 The Bender Gestalt Test of Visual Motor Integration

The original Bender Gestalt, called the Visual Motor Gestalt Test was first published by Lauretta Bender in 1938. She based it on the work of Max Wertheimer, one of the originators of the Gestalt School of Psychology who. In 1923, he developed a series of geometric designs to illustrate the perceptual principles of Gestalt Psychology and measuring human perception (Bender, 1938).

Bender selected nine of Wertheimer's original designs and modified five of these in her development of the Bender Gestalt Test as a way of "exploring deviations in the maturational process in perceptual motor functions..." (Tolor & Schulberg, 1962, p. 202). Bender used the designs in a test to differentiate brain-injured from non-brain-injured adults and to detect signs of emotional disturbance (Salvia & Ysseldyke, 1988).

Since the publication of her monograph, Bender's test has come into widespread use as a clinical instrument. It has been used to estimate maturation, intelligence, psychological disturbance, the effects of brain injury and to follow the effects of convulsive therapy (Pascal & Suttell, 1951)

The Bender Gestalt does not constitute a single instrument. There are many versions of the test with varying methods of administration, scoring and interpretation (Reichenberg & Rafael, 1992). The test was republished by Pascal and Suttell in 1951, and revised for persons over seven years of age by Hutt and Briskin in 1960. A Bender Visual Motor Gestalt Test specifically for children aged seven to eleven years was published by Clawson in 1962 (Tolor & Schulberg, 1962).

Bender herself did not provide an objective scoring system for the test. However, since her early work, numerous attempts have been made to develop a reliable scoring system. Billingslea (1948) was the first to publish a scoring system for the test. Additional scoring systems were developed by Gobetz (1953), Keller (1955), Kitay (1950) Peek and Quast (cited in Pascal and Suttell, 1951) and Stewart and Cunningham (1958). These scoring systems were all designed for use with adult psychiatric patients or with retarded and institutionalised children. They were not intended for use with younger children of adequate intellectual capacities (Koppitz, 1963).

2.3.5 The Koppitz System

The most recent Bender Gestalt Test developed specifically for young children was developed by Elizabeth Koppitz in 1964 (Savage, 1968). She provided a precise scoring method based on norms collected from children aged five to ten years. This methodology is the one most commonly used in school settings (Fuller, Awadh & Vance, 1998) and, for this reason, was the methodology used to evaluate the Bender protocols in this research.

The Koppitz Developmental Bender Scoring System is divided into two sections:

(1) developmental age scoring
(2) emotional assessment

The developmental scoring system consists of 30 scoring items; each item is scored 0 or 1 depending upon whether an error occurs. There are four types of errors: distortions of shape; rotation; integration difficulties and perseveration. The number of errors scored for each of the nine drawings is summed to yield the total error score. The total number of errors a child makes is then compared to the norms for the child's age. The higher the raw score, the poorer the performance (Fuller et al 1998, Koppitz 1963, Tolor & Brannigan 1962).

Each scoring item in the Developmental Scoring System was validated against first and second grade achievement as measured by the Metropolitan Achievement Test. Only those items were included which were able to differentiate statistically between above and below average students at the five per cent level, in either the first or second grades, or which demonstrated a strong trend, i.e. significant at the 10 per cent level, in both the first and second grades (Koppitz, 1963).

Koppitz proposed twelve emotional indicators, which were used to evaluate the emotional stability of the child via visual-motor performance. The emotional indicator scoring system is, however, infrequently used in school settings (Fuller et al, 1998).

2.3.6 Norms

Koppitz initially normed the test in 1963, and re-normed it in 1974. This latter sample consisted of 975 North American elementary school children ranging in age from 5 to 11 years of age. Adequate racial and demographic representation was not obtained, nor was the sample stratified according to socio-economic status (Fuller et al ,1998). The sample sizes for half year interval age groups in both the 1963 and 1974 samples are unevenly distributed (Salvia & Ysseldyke, 1988). Fuller et al (1998) further note that the standard deviations after age 8 years, 6 months were about equal to the mean for the 1974 norms.

2.3.7 *Inter-rater reliability*

Koppitz (1975) summarised 23 studies of inter-rater reliability for her scoring system. Inter rater reliability ranged form .79 to .99 with 81 percent exceeding .89. The 1975 manual also summarised the results of nine test-retest reliability studies with normal elementary school children. Reliability coefficients ranged from .50 to .90 (Koppitz, 1963). Salvia and Ysseldyke (1988) note however that five of the nine reliability studies that Koppitz reports on are on kindergarten children only, and only one of twenty five reported coefficients exceeding the standard of .90 recommended for tests used to make important decisions.

2.3.8 Predictive validity

Mckay and Neale (1985) found that results from the Bender Gestalt administered to first grade children were not good predictors of reading ability at the end of Grade 2. Similarly, Knoff et al (1986) found that the Bender Gestalt scores of gifted elementary school children did not account for a significant amount of achievement test variance.

Nielson and Sapp (1991) reported Bender Gestalt scores predicted achievement for a sample of low birth weight children. Goldstein & Britt's (1994) study demonstrated that various tests of visual motor integration (including the Bender Gestalt), showed no relationship between visual-motor co-ordination and achievement. A study by Shapiro and Simpson (1995) found no significant relationship between Koppitz errors vs. specific academic skills (e.g. reading and math). However, Bender performance was associated with cognitive abilities reflected in WISC-R subtests that represented Bannatyne's (1971) spatial perception factor (Block Design, Object Assembly, and Picture Completion) and field dependence - field independence (Kaufman, 1979).

Little evidence exists supporting the construct validity of the Bender Gestalt. Fifty four criterion-related measures of school achievement are reported in the manual. According to the findings of various studies reported in the manual, the Bender Gestalt is not a good predictor of academic achievement (Fuller et al 1998).

Buckley (1978, p. 336) proposes that "we have no conclusive body of proof that the instrument can be used to predict school achievement, neurological impairment, or emotional problems at a statistically acceptable level." This of course begs the question as to why the Bender Gestalt is routinely included in test batteries, and why conclusions as to a child's visual-perceptual functioning are made on the basis of the results yielded from this instrument. The published research also appears to lack recent inter-rater reliability studies, and therefore necessitates research in this area.

2.4 Assessment Of Emotional Functioning Based On The Human Figure Drawing

Drawing techniques involving human figures have constituted part of a psychologist's battery for most of the 20th Century (Yama, 1990) and have served as the foundation for many different interpretations of the drawer's personality (Kamphaus & Pleiss, 1991). Two main approaches to the use of drawings have been pursued: the use of the drawing for cognitive purposes on the one hand and for psycho-emotional purposes on the other. The development of the human drawing as a measure of intellectual functioning has been tackled by various researchers: Burt (1921); Fay (1923); Goodenough (1926); Harris, (1963) (cited in Yama, 1990). The most important work in the field was conducted by Goodenough who developed her Draw-A-Man Test in 1926. This tool was devised as a measure of intellectual functioning and was based on the assumption that certain aspects of drawing performance correlate with children's mental age and thus can be used as a measure of intelligence (Goodenough, 1926, Harris, 1963, Naglieri, 1988). Harris revised the test in 1963, extending it to be valid with older children and including more criteria in the test to improve its reliability and validity (Harris 1963, Koppitz 1968, Mortensen 1991).

The Human Figure Drawing as a measure of emotional adjustment and a description of personality functioning developed as a separate endeavour in the early part of this century. Lewis (1928) was the first to systematically describe personality on the basis of figure drawings (Mortensen, 1991). In 1949 Machover published a comprehensive analysis of the HFD grounded in psychoanalytic theory and the projective hypothesis. Her work has had a significant influence on how children's drawings are interpreted (Cummings, 1986, cited in McNeish & Naglieri, 1993).

2.4.1 The Human Figure Drawing as a projective technique

Drawings are considered to serve as projective techniques because they confront the child with an unstructured situation. This lack of structure means that the child has to *make meaning of the task by drawing on his or her own experience.* The ambiguity lies in the fact that there is very little direction from the examiner. The child is not told what to draw and can respond with any variation of size, content and placement for each drawing (Cox, 1993).

He/she is therefore highly likely to respond with themes from his/her own life experience, reflecting his/her personal needs (Gabel et al, 1986), with the result that the drawing is considered to provide valuable information as to the inner world, feelings and personality characteristics of the child. These may not be directly obtainable from verbal or other forms of communication (Rudenberg, Jansen & Fridjhon, 1998).

The use of projective techniques (Thematic Apperception Test, HFD) is based on the assumption that the individual is driven by psychological forces blocked from consciousness. The child reveals these unconscious conflicts by projecting his/her characteristic modes of response, thought processes, needs, anxieties and conflicts into the task.[3] (Anastasi 1982, Sabatino Fuller & Altizer, 1998).

Because projective techniques represent disguised testing procedures, as the individual is rarely aware of the type of psychological interpretation that will be made of his/her responses, they are considered to be particularly sensitive to the uncovering of unconscious or latent aspects of the personality (Anastasi, 1982).

2.4.2 The work of Machover

The use of the human figure drawing as a projective measure of personality functioning was first formally developed by Machover who published The Draw-A-Person Test in 1949. Her test was based on the assumption that there is an influential correspondence between the drawer and his/her drawing.

---

[3] The term projection is also employed in a much broader and encompassing fashion, and is commonly regarded as the general tendency to externalize aspects of the self. (Rabin, 1986).

Machover proposed that "the human figure drawing by an individual who is directed to 'draw a person' relates to the impulses, anxieties, conflicts, and compensations characteristic of that individual. In some sense, the figure drawn *is* the person and the paper corresponds to the environment" (Machover, 1949, p. 35).

As stated, Machover's work emerged from psychoanalytic theory and her theories drew on the "projective hypothesis" based on the defence mechanism of projection. This process is subject to a number of different interpretations. Machover used it in the sense that "specific impulses, wishes, aspects of the self, or internal objects are imagined to be located in some object external to oneself" (Rycroft, 1968:125). In other words, the individual attributes his or her own weaknesses or unacceptable desires to other people, thus decreasing feelings of anxiety that are associated with these desires (Sabatino, Fuller & Altizer, 1998).

Machover regarded distortions of the drawn figure to be symbolic representations of inadequacies in the drawer's self-image (Machover, 1949). According to her, drawings provide both direct and indirect expression to the drawer's basic needs and conflicts, which makes it possible to see the dominating defence mechanisms (Mortensen, 1991).

Machover's work has inspired an extensive body of research over the past forty years and numerous studies have been conducted to examine the reliability and validity of her assertions (Bruening, Wagner & Johnson, 1997, cited in Mortensen, 1991). In general, her formulations and inferences have come under considerable attack. The most comprehensive criticisms have come from Roback (1968) and Swenson (1957). These authors found little support for Machover's hypothesis concerning the meaning of the human figure drawings, and argued that the DAP was "of doubtful value in clinical work" (Swenson, 1957, p. 460-461). However, Swenson (1957) suggests that the test can be tentatively used as a rough screening device and a gross indicator of "level of adjustment" (Swenson, 1957, p. 650).

Other research (Cassel et al 1958, Strumpfer, 1963) has found relatively high inter-rater correlation's on human figure drawing protocols. Cassel et al (1958) reported that after three training sessions the analyses of judges as to the presence or absence of emotional signs correlated with each other above .90. Strumpfer assessed inter-rater correlation's on six different measures of general aspects of figure drawings such as overall quality, adjustment, sexual differentiation, maturity and body image disturbance. His inter-rater correlation's ranged from .79 to .97. Swenson (1986:21) suggests therefore that with a certain amount of training and/or explicit instructions, judges are able to rate figure drawings with "satisfactory reliability."

It appears that despite vociferous criticism levelled at Machover's Draw a Person Test, researchers continued, and still continue (Naglieri, 1988), to develop further methods of gaining insight into child and adult emotional functioning through the projective use of drawings. As an alternative to the interpretation of individual signs, a global assessment of emotional assessment was proposed by Koppitz (McNeish & Naglieri, 1993). This is discussed below.

2.4.3 The contribution of Elizabeth Koppitz

Some of the most influential work in use of the human figure drawing as a projective technique has come from Koppitz who, in 1968, proposed a method of analysing human figure drawings using specific signs as emotional indicators (Rudenberg et al, 1998). Her method involved counting the number of times an emotional indicator considered to be associated with disturbance appeared in the child's drawing and comparing this number to that found in the drawings of non-disturbed children (McNeish & Naglieri, 1993) Koppitz departed from the psychoanalytic theory employed by Machover and based her research on Sullivan's Interpersonal Relationship Theory. This was in line with the general trend in the 1960's to shift towards Ego Psychology and the analysis of conscious processes (Koppitz, 1968). Instead of focusing on the child's unconscious needs, defence mechanisms and psycho-sexual development, Koppitz emphasised rather the child's attitude toward him/herself, the most important persons in his/her life, and his/her attitude to current problems and conflicts.

She did not view the child's drawings as an expression of his/her basic and lasting personality traits, nor as the representation of his or her actual physical appearance, but rather as an expression of the child's current state of mind and attitudes (determined by both development and social and emotional conditions at any given moment). Koppitz described thirty specific traits or indicators in drawings. In order to be included as an indicator a trait had to be:

(1) clinically valid, i.e. able to differentiate between children with and without emotional problems.
(2) found in less than 16% of the children at any given age
(3) neither a function of age nor of maturity (Koppitz, 1968, Mortensen, 1991).

Koppitz operated with three different types of emotional indicators (Koppitz, 1968, Mortensen, 1991)

1) qualitative aspects such as poor integration, asymmetries and shading
2) presence of unexpected traits
3) absence of expected traits at certain ages

Koppitz suggested that the presence of two or more emotional indicators in a HFD was related to emotional disorder in children (Porteous, 1996). She compared the emotional indicators in the human figure drawings of two groups of children aged between 5 and 12 years.

Group one comprised of seventy six children attending ordinary schools and considered by their teachers to be well-adjusted. Group two comprised seventy six children who were patients at a child guidance clinic. Whereas the "normal" children produced a combined total of two emotional indicators, the clinic children produced a combined total of 166. Koppitz conducted further research into the validity of specific items. She consistently found aggressive children drawing items such as teeth; asymmetrical limbs; long arms; big hands, while shy children omitted the mouth more often than did aggressive children.

As with Machover's test, Koppitz's criteria have been the subject of extensive research. Support for their reliability and validity as indicators of emotional and behavioural adjustment have been mixed. Research has on the whole focused on the interpretation, validity and reliability of the various signs i.e. shading, big heads etc. The research in the area appears to be equivocal (Koppitz 1968, Mortensen 1991, Cox 1993). Research conducted by Tharinger and Stark (1990) found that the thirty emotional indicators were not able to distinguish between emotionally disturbed children and non-disturbed controls.

Authors such as Kahill (1984) and Motto, Little and Tobin (1993) have contended that structural and content variables are invalid indicators of maladjustment. However, other researchers have maintained that human figure drawings are useful tools in assessing personality as well as cognitive variables (Holtzman, 1993, Naglieri, 1993, cited in Bruening et al, 1997).

CHAPTER THREE: RATIONALE AND AIMS OF THE STUDY

3.1 Rationale For The Study

3.1.1 The role of the clinician in the assessment of children

At present, children are referred for assessment when they experience developmental delays, learning problems in the school environment, do not achieve academically or exhibit a variety of behavioural and emotional problems (Gabel et al, 1986). Assessment is a complex process and assessment data are used to make important decisions about individuals. The psychologist is in a unique position vis a vis the child he/she undertakes to assess. The sense the clinician makes of the child's difficulty and the referral he/she makes, sets the stage for certain interventions to take place. If the nature of the difficulty is effectively identified, there is a far greater chance for effective remediation. If effective remediation occurs, it is presumed that the child's life opportunities are likely to be significantly broadened. If, however, the child is insufficiently or inaccurately assessed his/her life opportunities can be adversely affected (Salvia & Ysseldyke, 1988).

Given the gravity of an assessment and the weightiness of the responsibility placed in the hands of the clinician, it is hoped that "diagnoses are made on the basis of thorough analysis of data rather than subjective speculation" (Ysseldyke & Salvia, 1988: 280). Notwithstanding, it is apparent from the different impressions of the same child obtained from the same clinician, as well as the variance observed in results that, far from being entirely objective, the assessment process involves a strong subjective component (Mattarazo 1990, cited in Vance, 1998). Indeed, psychological tests - such as projective tests - are notorious for the high degree of clinical judgement that they require.

As Swenson (1968) notes, human figure drawings rely heavily on clinical judgement and as a consequence are dependent on the interpreter's intuition. When recommendations are made on the basis of tools where considerable variance is evident, a significant concern as to the validity of the results can be expected. In terms of this, establishing the intra and inter-rater reliability of such tests is considered by the researcher to effectively establish an important aspect of their clinical utility.

3.1.2 The Bender Gestalt and Human Figure Drawing

Despite the controversy surrounding their continued use, projective tests such as the HFD are among the most frequently administered. However, (Sabatino, Fuller & Altizer, 1998) argue that this type of tool, due to the subjective nature of its scoring, generates low reliability, making interpretations difficult. Furthermore, different clinicians using the same scoring systems derive scores which may vary greatly (Kline 1993, Sabatino et al, 1998). Similarly, the Bender Gestalt is by far the most frequently used test of visual-perceptual functioning (Gabel et al, 1986). However, only equivocal evidence has been produced to support the Bender's construct validity and predictive reliability (Buckley 1978, Vance, Fuller & Lester, 1986, cited in Vance 1998).

Meehl (1954, cited in Salvia & Ysseldyke, 1988) numerates five psychometric standards to which assessment instruments are expected to adhere. One of these, test reliability, i.e. the consistency of scores from the same test over various administrations, is the foundation of this report. A survey of the literature indicates that while there is abundant research into the construct and content validity as well as test-retest and predictive reliability of the Bender Gestalt and HFD in publication, there appears to be a dearth of literature examining specifically the inter-rater reliability of the tools. This is puzzling considering the pivotal role played by the clinician in determining the nature of the child's difficulty and in making an appropriate referral.

3.1.3 The Bender Gestalt and Human Figure Drawing in the South African context

Following the integration of previously disadvantaged learners into a unified education system, a large number of children are presenting with learning difficulties. Many are eventually referred to public sector psychological services, as private assessment is costly and invariably beyond the financial means of the vast majority of South Africans. Unfortunately, public sector services are generally cash-strapped and understaffed, with the result that it is necessary to conduct psychological assessments that are time and cost efficient and to proceed in such a manner that does not disadvantage English second language speakers.

This means, *de facto,* that only a few of the many available assessment tools are being used. Of these the Bender Gestalt and HFD are routinely being included in test batteries to assess children presenting with learning and/or emotional maladjustment. In fact, in many instances, the Bender Gestalt and HFD comprise two out of only three or four assessment tools used. Indeed, the Bender Gestalt and HFD are considered by some clinicians to be particularly applicable to the South African context, as they are quick and easy to administer, score and evaluate. In addition, the HFD is a task that appeals to children of all cultural backgrounds and, for the above reasons, has been extensively applied in numerous and diverse cultural settings (Richter, Grisel and Wortley 1988). It is evident that their widespread use in the South African context increases the importance of establishing their reliability.

3.1.4 The most common uses of the HFD and Bender Gestalt in South Africa

Based on experience in a Gauteng clinic and a private therapy and assessment centre over the past two years, this researcher observes that some of the most frequently made referrals following an assessment are:

(1) to a neurologist or paediatrician for the further investigation of possible neurological deficits such as attention difficulties;
(2) to a speech and hearing therapist for further assessment and/or remediation of speech and hearing difficulties;
(3) to an occupational or remedial therapist for the further assessment and/or remediation of visual-motor and visual-perceptual difficulties;
(4) to a play therapist/child psychotherapist to assist the child with underlying or apparent emotional difficulties.

While the researcher believes that referrals (1) and (2) are unlikely to be established through the sole use of the Bender Gestalt and HFD[4], referrals (3) and (4) are. Part of establishing these tests reliability and clinical efficacy would, therefore, have to include the clinicians ability to distinguish between visual-perceptual and/or emotional difficulties when scoring these tests.

3.1.5 Conclusion

Assessment is at a crucial stage of its development. Attacks on assessment procedures, practices and instruments are common place. Some educational professionals find the results of tests irrelevant and unresponsive to the clinical or educational needs of children. They find that assessments frequently bear no tangible relation to remediation planning (Yama, 1990). "This problem has probably developed because psychologists have in some cases, formed a misguided and self-destructive loyalty to the standard battery of tests (DAP, Bender etc.)" (Vance & Awadh, 1998, p. 12). This does not, however, appear to have deterred the enthusiasm with which clinicians abroad and in South Africa routinely administer and draw inferences from Bender Gestalt and HFD.

It does, however, appear that the Bender Gestalt and HFD may be particularly attractive in South Africa where non-verbal, non-biased, easy-to-administer and cost-effective measures are required to assess large numbers of children. It is further evident that clinicians routinely administer these tools and frequently base their understanding of the child's difficulty, in part, on the basis of the results of these tools, i.e. a clinician may "categorise" the child to be, for example, "emotionally disturbed" on the basis of the results obtained from minimal test data. It is hoped that all children will have an equal opportunity of having the nature of their particular difficulty explicated by the clinician to whom they are referred.

[4] Referral (1) because it is also necessary to observe the child interacting with his/her environment, and to observe his/her ability to attend. Referral (2) because The Bender Gestalt and the HFD do not necessarily require any verbalisation on the part of the child.

Because of the situation described above, this research project primarily aimed to investigate whether the Bender Gestalt and Human Figure Drawing are being usefully applied in clinical practice. This was based upon establishing whether two independent clinicians were able to distinguish consistently between the child experiencing visual-perceptual difficulties, and the child experiencing emotional difficulties, (or to motivate as to the presence of both difficulties) and on the basis of this to make reliable and consistent analyses and formulations of children's difficulties.

3.2 Aims Of The Study

The specific aims of the study were to establish the following:

1) The level of agreement between two skilled, independent raters scoring 52 children's protocols of both the Bender Gestalt and HFD using the Koppitz scoring systems (inter-rater agreement, first rating session).

2) The level of agreement between the two raters' scores at a repeated rating after a three week interval (inter-rater agreement from the first to the second rating sessions).

3) The level of agreement between the raters' "diagnosis" [5] of the child's difficulty i.e. visual-perceptual; emotional; experiencing both difficulties; experiencing neither difficulty, on the basis of an analysis of the results of the two measures (inter-rater agreement in terms of diagnosis, session one)

4) The level of agreement between the raters' diagnosis following a three week interval (inter-rater agreement in terms of diagnosis, from session one to session two)

5) The level of agreement between each individual rater's scores and diagnosis from the first and second rating sessions (intra-rater agreement, in terms of both scores and diagnoses following a three week interval)

It is intended that this research will contribute to the ongoing body of research examining the reliability of these two psychological assessment tools and will provide valuable information regarding whether their current usage in clinical practice is warranted.

[5] Although the researcher considers the term "diagnosis" to be more appropriate in a medical context, the term is used throughout for the sake of brevity (rather than e.g. clinical formulation)

CHAPTER FOUR: METHOD

4.1 Subjects

The subjects of the research were two independent raters who were chosen on the basis of their experience in clinical assessment and their particular expertise in examining the Bender Gestalt and HFD tools. Both raters (one male, one female) are research psychologists and fully trained psychometrists. They each have at least two years clinical experience in addition to their clinical training.

4.2 Sample

In this study the sample consisted of the 204 ratings generated by the two raters from the scoring of 52 children's HFD and Bender Gestalt protocols[6] on two separate occasions. The protocols were selected from children who sought assistance for assumed learning and/or emotional difficulties at a private therapy and assessment centre in the Johannesburg North area.

4.3 Sampling procedure

The protocols were randomly selected on the basis of a computer generated random numbers table. The computer generated random file numbers from within the range of file numbers used by the clinic.

4.4 Age at time of test

Note that ages are reported in decimal years, that is a mean age of 8.86 years is 8 years and 86/100. This statistic was calculated with the original sample size of 54. However, two protocols had to be discarded as they lacked the appropriate demographic data, i.e. age and gender. Henceforth, the sample size is referred to as 52.

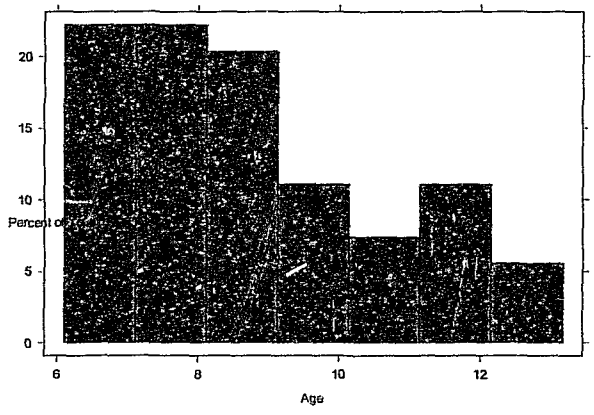
---

[6] Permission to use this sample was obtained from the trustee of the clinic

Table 1 Summary statistics for age at test

N	Mean	Std Dev	Minimum	Maximum
54	8.864704	1.864823	6.165640	13.084188

Figure 1 Age distribution



4.5 Distribution of rater by test

The protocols were rated at two different rating sessions and are referred to as Tests 1 and 2. Two different raters scored and evaluated the protocols into the four categories which are detailed further in this chapter. The raters are referred by the first letter of their first names: "M" and "R". The following table reflects the number of ratings per test per rater.

Table 2 Distribution of rater by test

Frequency Percent Row Pct Col Pct	Test		
	1	2	Total
M	52 25.00 50.00 50.00	52 25.00 50.00 50.00	104 50.00
R	52 25.00 50.00 50.00	52 25.00 50.00 50.00	104 50.00
Total	104 50.00	104 50.00	208 100.00

4.6 Instruments

4.6.1 Bender Gestalt Test of Visual Motor Integration

This test was selected in order to establish:

(1) The reliability of the raters' scoring of the children's visual-motor-perceptual functioning.

(2) The reliability of the raters' diagnoses of the children's difficulties.

The Bender consists of nine geometric figures, drawn in black on a 10.16 cm by 15.24 cm white card. The test is administered individually and is not timed. The child is asked to draw all nine designs (one at a time) on a blank sheet of A4 white paper. A child completes copying the designs in an average of five minutes (Fuller, Awadh & Vance, 1998).

The developmental scoring system consists of thirty scoring items; each item is scored 0 or 1 depending upon whether an error occurs, i.e. if an error is noted the protocol receives an error score for that item. There are four types of errors: Distortions of shape, rotation, integration and perseveration. The number of errors scored for each of the nine drawings is summed to yield the total error score. The total number of errors a child makes is then compared to the norm for the child's age, which is converted into a developmental age equivalence and represented as an age range. The higher the raw score, the poorer the performance (Koppitz 1963, Tolor & Brannigan 1968, Fuller Awadh & Vance 1998). According to Koppitz (1963) a child scoring approximately one year below his/her actual age, is said to be experiencing problems in the visual-perceptual domain.

4.6.2 The Human Figure Drawing

This test was selected in order to establish:

(1) The reliability of the raters' scoring of the children's emotional functioning.

(2) The reliability of the raters' diagnosis of the children's difficulties.

As with the Bender Gestalt, this test had already been administered and the data collected. At the time of testing (which occurred between January 1996 and October 1997), each child was provided with pencil and eraser and with one sheet of A4 paper on which to complete the drawing. He or she would all have been asked to (1) "draw a picture of a whole person and not to make a stick figure or cartoon". Originally these drawings were analysed on a qualitative level by the clinicians who assessed the children, and the Koppitz model was not applied at that time.

4.7 Procedures

Each HFD and Bender Gestalt was assigned a code number to ensure the raters remained unaware of the children's identity. Two rating sessions were conducted with a three week interval in between.

4.7.1 Session 1

Each rater was presented with three scoring sheets for each child. One sheet was for scoring the HFD, one sheet was for scoring the Bender Gestalt, and one sheet was for recording the diagnosis of the child's difficulty, which was to be based on the results of the HFD and Bender Gestalt scores. The tests were presented to the raters by child, i.e. each rater would score both the HFD (number of emotional indicators) and the Bender Gestalt (establish the developmental age equivalence) of one child, and then fill out a diagnostic protocol for that child, before moving on to the HFD, Bender and diagnostic protocol for the next child. The raters were informed of the age and sex of each child.

Once the scoring was completed, on basis of their analyses, the raters were requested to categorise the child on their diagnostic protocols into one of the following categories:

(1) Primarily exhibiting emotional maladjustment - this diagnosis was to be made on the basis of a high HFD score, i.e. two or more emotional indicators, AND an age appropriate Bender score, i.e. no more than a year behind chronological age.

(2) Primarily exhibiting visual-perceptual difficulties - this diagnosis was to be made on the basis of a high Bender score, i.e. at least one year behind chronological age, AND a low HFD score, i.e. no more than one emotional indicator.

(3) Exhibiting both difficulties - this diagnosis was to be made on the basis of a high Bender score i.e. at least one year behind chronological age, AND a high HFD score, i.e. at least two emotional indicators.

(4) Exhibiting neither difficulty - this diagnosis was to be made on the basis of an age appropriate Bender score i.e. no more than a year behind chronological age, AND a low HFD score, i.e. no more than one emotional indicator.

As Koppitz suggested, a child was considered to be experiencing emotional difficulties if he or she presented with two or more emotional indicators. As far as the Bender Gestalt was concerned, a child was considered to be experiencing visual-perceptual difficulties if his/her developmental age equivalence on the Bender Gestalt fell at a minimum of one year below chronological age.

The rating session was 6 hours in duration. The raters took frequent breaks as well as an hour's meal break to ensure that their analyses were not affected by fatigue. Scoring manuals were supplied for both tests, as well as rulers and protractors to use as required (See Appendices A and B for descriptions of the Bender Gestalt and HFD scoring criteria).

4.7.2 Session 2

Following a three week interval, the same procedures as at session 1 were repeated. This session was conducted in order to establish the intra-rater reliability of the test scores and diagnoses over a period of time. At session 2, however, the protocols were presented in a different (also random) order to Session 1.

4.8 Specific Hypotheses

(1a) H_0 : No significant difference will be evident between the raters' scoring of the Bender Gestalt protocols versus H_1 : that there will be a significant difference between the raters' scoring of the protocols.

(1b) H_0 : No significant difference will be evident between the raters' scoring of the HFD protocols versus H_1 : that there will be a significant difference between the raters' scoring of the protocols.

(2) H_0 : No significant difference will be evident between the two raters' diagnosis of the child's difficulty versus H_1 : that a significant difference will be evident .

(3a) H_0 : No significant difference will be evident between the two raters' scoring of the Bender Gestalt protocols following a three week interval versus H_1 : that a significant difference will be evident.

(3b) H_0 : No significant difference will be evident between the two raters' scoring of the HFD protocols following a three week interval versus H_1 : that a significant difference will be evident.

(4) H_0 : No significant difference will be evident between the two raters' diagnosis of the child's difficulty following a three week interval versus H_1 : that a significant difference will be evident.

4.9 Statistical Procedures

The statistical analyses that were applied were all 2-tailed tests, i.e. the researcher was interested in any significant change from that referred to by the null hypothesis, i.e. whether the level of agreement is higher or lower than that postulated by the null hypothesis.

4.9.1 Summary statistics

These include the number of observations, the means, the standard deviations and the minimum and maximum of the scales in question. They are intended to provide an overview of the data rather than answering any specific hypotheses.

4.9.2 Analysis of Variance

An Analysis of variance (a generalisation of the traditional *t*-Test) was computed for each hypothesis. This method was applied as it is designed to detect differences in means (in this instance the difference in the means between the scores arrived at by M and R in terms of the Bender Gestalt and HFD). The actual computations were performed using the SAS statistical software package.

4.9.3 Pearson's Correlation Coefficient

Because each protocol provided a numeric score, these were also evaluated according to traditional correlation techniques (Pearson's Correlation Coefficient). This method analyses the strength of the linear relationship between variables.

Correlation's between the first and second session scores for each rater and each score were computed to determine the level of agreement between the first and second sessions for each rater, thereby enabling the assessment of the degree of intra-rater reliability. Correlations between the scores for each rater irrespective of session were also computed, enabling an assessment of the inter-rater reliability of the raters.

In both cases correlation's significantly greater than zero were taken to signify a linear relationship, i.e. the magnitude of the correlation (irrespective of whether it was positive or negative) was considered to be a direct measure of the reliability of the relationship.

4.9.4 Tests of Association/Agreement

Tests of association were applied in terms of the diagnosis of the children's difficulties. These were chosen because the diagnoses fell on a nominal scale, i.e. there are four nominal classes (emotional; visual-perceptual; both; neither) and the aim was to compare the diagnoses between two different classifications, i.e. Test 1 vs. Test 2 or Rater M vs. Rater R.

4.9.4.1 Chi Square

On the basis of the results of each child's Bender Gestalt and HFD protocol, he/she was assigned to one of the above four categories by the rater. In order to establish the level of intra and inter-rater agreement, it was necessary to compute the level of association with respect to the raters' assignment to one of the four classes - on the one hand between Tests 1 and 2 (intra-rater agreement) and, on the other hand, between the two raters (inter-rater agreement).

It was first necessary to construct a contingency table which cross-tabulated the raters' diagnoses, i.e. counted the frequency of each combination of diagnosis. This information is presented on the contingency table. The next step was to analyse the strength association between the column and row axes. The Chi Square statistic was then computed from the contingency table. This test statistic measures the deviations between the observed and expected values in each cell of the table. The larger the difference between the observed and expected values, the larger the Chi Square value was and the more evidence there was to support the rejection of the null hypothesis. The Chi Square statistic was chosen in this instance as it is widely used for this type of comparison especially as it is a method that relies on a relatively large sample (Fleiss, 1981).

The Likelihood Ratio Chi Square was also computed. This was used as it makes allowance for instances where the expected value is smaller than one, a case that can lead to incorrect conclusions when the traditional Chi-Square statistic is applied (ibid, 1981).

4.9.4.2 Fisher's Exact Test

This statistic is a non-parametric calculation of the probability of observing a more extreme arrangement of the table than was the case. It was used as it is an exact calculation, rather than an approximation. It does not make any assumptions about an underlying normal distribution or large sample approximation, as does the Chi-Square statistic. The complexity of the calculation grows exponentially with the sample size and so this method is only readily computable on modest samples (Fleiss, 1981).

Since the sample was of a relatively modest size, it was not clear which of the methods would be the most appropriate. It was, therefore, decided to use both Chi Square and Fisher's Exact Test.

4.9.4.3 Cohen's Kappa

Cohen's Kappa was computed because it was designed specifically to measure agreement between ratings on a nominal scale. Following Cohen, any value greater than 0.4 was considered fair agreement beyond chance (Stuart & Ord 1973, Rudenberg et al, 1998).

CHAPTER FIVE: RESULTS

5.1 Inter-Rater Reliability - Scoring Of The Bender Gestalt Test

5.1.2 Summary of the Bender Gestalt Test results - inter rater comparison (H_0 1a)

Summary statistics were computed and are presented below. From Table 3 it is evident that there was a difference in means between the raters. This difference was formally tested by the ANOVA procedure which follows.

Table 3 Summary statistics – Bender Gestalt Test - inter-rater comparison

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
M	103	8.4656	1.5311	871.9583	5.2083	11.4583
R	104	7.8970	1.6877	821.2917	5.0417	11.4583

5.1.3 Analysis of Variance of the Bender Gestalt Test

A linear model which included all of the hypothesised relationships between the Bender score, the test session and the rater, was fitted to the data in order to establish whether there were any significant differences in terms of the scoring of the Bender Test, (1) between the raters (inter-rater reliability), and (2) between the first and second scoring sessions (intra-rater reliability). The dependent variable in this instance was the Bender Age less the Actual Age at the time of the test (referred to as the BMA). A negative BMA implied that the child's Bender Age was lower than their actual age.

In order to establish the level of agreement between the two raters, the scores arrived at by each rater were compared to each other. For the Bender test, it was necessary to find the midpoint of the age range arrived at by the raters.

The ANOVA table for this model fit is displayed below. (An ANOVA or analysis of variance table is the summary of the variability taken into account by the model, along with a determination of whether this constituted a significant de e of variation).

Table 4 Analysis of variance for dependent variable: BMA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	26.9772611	8.9924204	3.10	0.0278
Error	203	589.0536655	2.9017422		
Corrected Total	206	616.0309266			

The R-Square value 0.043792 indicated that the model was of little predictive value.[7] The model was however significant at the 5% level, i.e. Pr > F is less than 0.05, suggesting that a significant factor was involved. The ANOVA table for each factor was examined to determine which factor was significant, i.e. whether the significance was at the level of the rater, at the level of differences between Tests 1 and 2, or at a level of the interaction between these two.

Table 5 Analysis of variance for factors modelling BMA

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Rater	1	16.6901682	16.6901682	5.75	0.0174
Test	1	5.2567072	5.2567072	1.81	0.1798
Rater*Test	1	4.9818338	4.9818338	1.72	0.1916

It is evident that there was a significant difference between the raters with regards to the Bender Test, i.e. Pr > F is once again less than 0.05 (inter-rater reliability). This was not true of the test, or of the rater/test interaction, i.e. whether the rater influenced the test, or vice versa, thereby indicating that it was necessary to further investigate the nature of the inter-rater rather than the intra-rater, reliability. The mean BMA for each rater was therefore established and the nature of the difference tabled below:

Table 6 Analysis of Variance - Comparison between BMA means per rater

	BMA		
Level of rater	N	Mean	SD
M	103	-0.41719580	1.48108220
R	104	-0.98507921	1.90959437

[7] The R-Square value does not appear on the table but is a calculation arrived at from the table

The difference between M and R's means indicated that R rated the tests almost seven months more negatively than did M (The actual difference in months is calculated by taking the difference between the means - .568 and multiplying this number by 12 months, i.e. 6.816 months). This ANOVA table therefore indicates a low level of agreement between the raters (low inter-rater agreement).

5.1.4 Correlation Analysis - Pearson's Correlation Coeffecient

Pearson's Correlation Coefficient was computed to examine the strength of the linear relationship between the raters

Table 7 Pearson's Correlation Coefficient – Bender Gestalt inter-rater comparison

Pearson Correlation Coefficients Prob > |R| under Ho: Rho=0* Number of Observations	M	R
M	1.00000 0.0 103	0.70349 0.0001 103
R	0.70349 0.0001 103	1.00000 0.0 104

* This refers to the probability that correlation of .70349 is observed given that the population correlation is actually zero.

The correlation between the raters of 0.70349 with a p-value of 0.0001 indicated a measurable and detectable linear relationship between M and R's scores. Therefore when, for example, the rater M scored a protocol with a high Bender score, the rater R also gave the protocol a relatively high score (although not necessarily the same score, as reflected by the 7 months difference on average between their scores). Similarly, when M scored a protocol with a low Bender score, so did R. As mentioned, the ANOVA results (Table 6) indicated a significant bias between the raters.

Although a linear relationship (as per the correlation analysis) was detected between the raters, it was necessary to analyse the computations from the ANOVA tables, in order to ascertain the actual level of inter-rater reliability. This was necessary as the ANOVA provided a comparison between means, where a measurable difference was observed. In this instance, therefore, the null hypothesis (1a) was rejected in favour of the alternative hypothesis, i.e. that a significant difference was evident between the raters in terms of their scoring of the Bender Gestalt protocols.

5.2 Intra-Rater Reliability Of The Scoring Of The Bender Gestalt Test (3a)

5.2.1 Summary statistics

Rating sessions 1 and 2 are referred to throughout as Tests 1 and 2. The analyses of M and R are considered together. The summary statistics show the means to be within one standard deviation of each other. This suggests that a significant difference between the Tests will not be evident in terms of the ANOVA.

Table 8 Summary statistics – Bender intra-rater comparison

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Test 1	103	8.3398	1.6632	859.00	5.2083	11.4583
Test 2	104	8.0216	1.5943	834.25	5.0417	11.4583

It is now necessary to refer to the ANOVA tables (4 and 5) as these provide information as to the nature of the relationship between Tests 1 and 2. It is evident from the ANOVA table that it is not possible to reject the null hypothesis, i.e. that there was no significant different between Tests 1 and 2 in this case (3a). This is taken as evidence that there is indeed a consistency between the testing sessions, i.e. intra-rater reliability was apparent.

5.2.2 Correlation analysis - Pearson's Correlation Coefficient

Pearson's correlation coefficient was computed to examine the strength of the linear relationship between the raters' scoring of the Bender Gestalt from Tests 1 to 2.

Table 9 Pearson's Correlation Coefficient – Bender intra-rater comparison

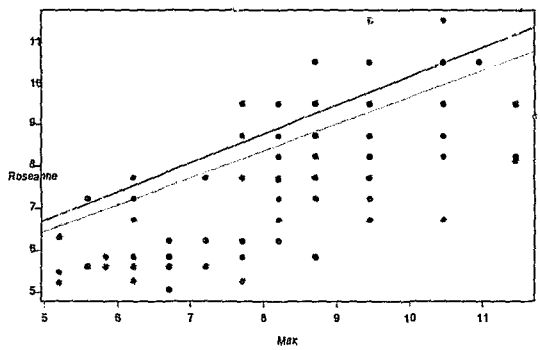
Pearson Correlation Coefficients Prob > |R| under Ho: Rho=0* Number of Observations	Test 1	Test 2
Test 1	1.00000 0.0 103	0.76187 0.0001 103
Test 2	0.76187 0.0001 103	1.00000 0.0 104

* This refers to the probability that correlation of .76187 is observed given that the population correlation is actually zero.

The correlation of 0.76 between Tests 1 and 2 indicates a measurable and detectable linear relationship.

In summary the ANOVA table indicated that a sufficiently high level of agreement between Tests 1 and 2 was evident. Further, the correlation analysis showed a reasonably strong relationship between Tests 1 and 2. It is therefore not possible to reject the null hypothesis in this instance (3a) i.e. that no significant difference was evident between the two raters' scoring of the Bender Gestalt protocols following a three week interval.

Figure 2 Bender correlation's across rater by test



The scatter diagram illustrates both Tests 1 and 2 as well as the differences between the raters. In these plots, the cyan points represent the first test, while the magenta represent the second. M's scores run along the horizontal axis, while R's run along the vertical axis. The values plotted are the actu l Bender Scores. This diagram illustrates the estimated half year difference in score between the two raters. A score of 7 by M translates to a score of 7.75 by R at the first test and a score of 8 at the second. The correlation is reflected in the slope of the diagonal lines. Since these lie roughly parallel to each other and are about 0.7 years apart. This confirms the biased relationship shown by the ANOVA and correlation analyses. The lines refer to the relationship between M and R's ratings at Tests 1 (magenta) and Test 2 (cyan).

5.3 Inter And Intra-Rater Reliability of The Scoring Of The Human Figure Drawing (HFD) (3b;1b)

A linear model that included the rater, the test, and the interaction between the rater and the test, was computed in order to predict the HFD sc re. This yielded the analysis of variance tabled below (since no significant factors were detected by the ANOVA, the summary statistics were not reproduced as they didn't further the argument in any appreciable manner).

5.3.1 Analysis of Variance

Table 10 Analysis of variance for dependent variable: HFD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.17307692	1.72435897	0.80	0.4927
Error	204	437.26923077	2.14347662		
Corrected Total	207	442.44230769			

This model has an R-Square value of 0.011692, indicating that as with the Bender Gestalt, the model explained very little of the observed difference. The model itself is not significant, i.e. Pr > F = 0.4927 is much larger than the 0.05 required for significance at the 5% level.

It is possible to infer that there was insufficient evidence to show a significant difference between raters. As with the Bender Test, the correlation analysis was computed, in order to gauge the strength of the linear relationship between the raters, and thereafter to determine the strength of the linear relationship between Tests 1 and 2.

5.3.2 Correlation analysis - Pearson's Correlation Coefficient

Table 11 Correlation analysis – HFD inter-rater comparison

Pearson Correlation Coefficients Prob > |R| under Ho: Rho=0*	M	R
M	1.00000 0.0	0.64790 0.0001
R	0.64790 0.0001	1.00000 0.0

* This refers to the probability that correlation of .64790 is observed given that the population correlation is actually zero.

A reasonably strong positive relationship between the raters was evident. Besides the means being within a standard deviation of one another, implying no mean difference between the raters, a correlation of 0.64790 was computed. This has a p-value of 0.0001, which implies that the correlation was significantly different from zero, and hence that there was a measurable and detectable relationship between the two raters in as far as the HFD test was concerned.

5.4 Intra-Rater Reliability Of The Scoring Of The Human Figure Drawing (HFD)

5.4.1 Correlation analysis - Pearson's Correlation Coefficient

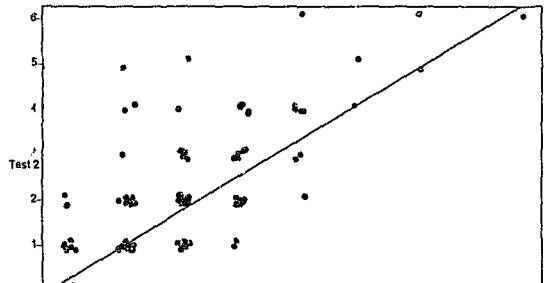
Table 12 Correlation analysis – HFD intra-rater comparison

Pearson Correlation Coefficients Prob > |R| under Ho: Rho=0*		
	Test 1	Test 2
Test 1	1.00000 0.0	0.73575 0.0001
Test 2	0.73575 0.0001	1.00000 0.0

* This refers to the probability that correlation of .73575 is observed given that the population correlation is actually zero.

The ANOVA table above indicates that there was no significant difference between the scores arrived at from Test 1 to Test 2. Also that there is a non-zero correlation (0.73575) between the HFD scores achieved during the two sessions. There is, therefore, insufficient evidence to reject the null hypothesis (3b) in this case, i.e. that no significant difference was evident between the two raters' scoring of the HFD protocols following a three week interval, in favour of the alternative hypothesis that a significant difference was evident.

Figure 3 HFD correlation's across test by rater



The different colour points on the scatter diagram are generally interspersed, which graphically represents the lack of significant differences found. We also see the level of agreement between the two tests and the two raters. The straight line superimposed on the plots represents a "linear relationship" The correlation measures how strongly the sample points adhere to this line. The fact that the dots are clumped together indicates a relatively high level of agreement between the two raters. However it is evident that M (magenta) displays less scatter than does R (Cyan), indicating greater consistency within himself than R within herself from Test 1 to Test 2.

5.5 Statistics For Intra-Rater Agreement Of The Diagnosis Of The Child's Difficulty (4)

In order to establish the strength of the relationship between the Test 1 and Test 2 in terms of the diagnosis of the children's difficulties, a number f analyses were conducted. Firstly a contingency table was constructed. This was followed by the Chi Square, Fisher's Exact and Cohen's Kappa statistics.

5.5.1 Contingency table

Table 13 Contingency table - comparison of analysis by test session

	Frequency Percent Row Pct Col Pct	Test 2				
		both	emotional	neither	visual-perceptual	Total
Test 1	both	4 3.85% 36.36% 15.38%	4 3.85% 36.36% 11.43%	0 0.00 0.00 0.00	3 2.88% 27.27% 18.75%	11 10.58%
	emotional	15 14.42% 40.54% 57.69%	15 14.42% 40.54% 42.86%	4 3.85% 10.81% 14.81%	3 2.88% 8.11% 18.75%	37* 35.58
	neither	1 0.96% 2.38% 3.85%	16 15.38% 38.10% 45.71%	22 21.15% 52.38% 81.48%	3 2.88% 7.14% 18.75%	42* 40.38%
	visual-perceptual	6 5.77% 42.86% 23.08%	0 0.00 0.00 0.00	1 0.96% 7.14% 3.70%	7 6.73% 50.00% 43.75%	14* 13.46%
	Total	26* 25.00%	35* 33.65%	27* 25.96%	16* 15.38%	104 100.00%

The following key will be utilised to refer to the four diagnostic categories:

"ED" category: A child considered by the rater to be experiencing primarily emotional difficulties

"VPD" category: A child considered by the rater to be experiencing primarily visual-perceptual difficulties.

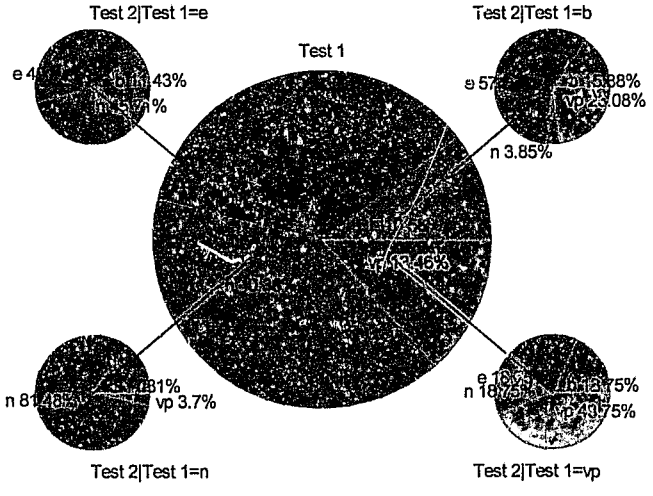
"BD" category: A child considered by the rater to be experiencing both difficulties.

"ND" category: A child considered by the rater to be experiencing neither difficulties.

5.5.2 Pie chart

The following pie chart graphically displays the difference between the ratings at the first and second test. The large centre pie represents the percentage of the sample analysed as Emotional (e), Neither (n), Both (b), or Visual-Perceptual (vp) at Test 1. The smaller pies around the edges represent the analyses at the second testing for those in the associated first test group. i.e. of the 35.58% of the sample analysed as emotional at the first test, 42.86% were analysed as emotional at Test 2. These figures correspond to those in the contingency table above.

Figure 4 Pie chart: Test 1 vs Test 2



From the contingency Table 13 it is evident that at Test 1 the raters assigned 11 diagnostic protocols to the category of experiencing BD[8]. This constituted 10.58 percent of the of total sample. At Test 2 they assigned 26* protocols to the category of BD (25 percent of the total sample). At Test 1, the raters placed 37* children in the ED category (35.58 percent). At Test 2, they assigned 35* to this category (33.65 percent). At Test 1, they diagnosed 42* children as falling into the ND category (40.38 percent), whereas this number fell to 27* at Test 2 (25.96). In the first session they categorised 14* children to the VPD category (13.46 percent), while this number rose to 16* at Test 2 (15.38 percent).

A closer analysis reveals the following: At the second rating session, 25 percent of the sample were classified by the raters as experiencing BD. Of this group, only 15.38 were classified with BD at Test 1. A total of 57.69 percent were assigned to the ED category and 23.08 into the VP category. This shows that the BD category expanded considerably and it appears that this sample was largely derived from protocols that had been assigned to the ED category at the Test 1.

Further analysis reveals the following: A total of 33.65 percent of the sample were rated as experiencing ED at the Test 2. Only 42.86 of this sample were classified with the same difficulty at Test 1. 45.71 percent were considered to experience ND at Test 1, 11.43 percent were derived from the BD category, and no children were placed in the VPD category. Therefore, the total number of children classified with primarily ED decreased slightly at Test 2.

At Test 2, 25.96 percent of the sample were classified as falling into the ND category. Of these 81.48 percent of this sample had also originally been classified into this category at the Test 1; 14.81 percent were derived from the ED category; 3.70 percent were derived from the VPD category and no children had previously been classified into the BD category at Test 1.

---

[8] The term "diagnostic protocol" is used only once. Hereafter the assignation to a diagnostic category, i.e. emotional; visual-perceptual; neither or both, is referred to merely as "protocol"

Therefore, although the total percentage of children classified to the ND category shrinks from 40.38 percent of the total sample to 25.96 of the sample at Test 2, it is evident that a large percentage of the same children (81.48 percent) were placed in the ND category at both tests. Although the ND category shrinks, it is evident that the highest level of agreement occurred in this category, perhaps indicating that the ND category is the most reliable.

At Test 2, 15.38 percent of the total sample of protocols were classified into primarily exhibiting a VPD. Of this sample, 43.75 percent had also been classified into the VPD category at Test 1; 18.75 percent of the sample of 15.38 had been classified into the BD category; 18.75 into the ED category and 18.75 into the ND category. This indicates that an equal percentage of the protocols which had been classified into the other three categories at Test 1 were re-classified by the raters into the visual-perceptual category at Test 2. Therefore the VPD category appears to expand slightly at Test 2, but less than 50 percent of the protocols that were classified into the VPD category at Test 1 were classified into the VPD category at Test 2. Again, this illustrates the lack of concordance between the raters' diagnoses from Test 1 to Test 2.

An analysis of the main diagonal on the contingency table illustrates the following: There was an agreement of 4 protocols at the level of the BD category between Test 1 and 2; an agreement of 15 at the ED category, between Tests 1 and 2; an agreement of 22 at the level of the ND category, and of 7 in the VPD category. This illustrates that the highest level of agreement between Tests 1 and 2 was in the ND category and the lowest level of agreement in the BD category.

5.5.3 Statistics for intra-rater agreement

Table 14 Statistics for intra-rater Agreement - Chi Square, Likelihood Ratio Chi-Square, Fisher's Exact Test, Cohen's Kappa

Statistic	DF	Value	Prob
Chi-Square	9	54.398	0.001
Likelihood Ratio Chi-Square	9	62.222	0.001
Fisher's Exact Test (2-Tail)			4.36E-10
Cohen's Kappa		0.2606322	
Observed Agreement		0.4615385	
Expected Agreement		0.2717271	
Normal Approximation		4.591531	2.20e-006

The Chi square statistic of 54.398 with 9 degree of freedom yielded a probability level of 0.001. This figure is much smaller than the significance level (0.05). This indicates that very little agreement was evident between Tests 1 and 2 in terms of the raters' diagnoses of the children's' difficulty. It is, therefore, necessary to reject the null hypothesis (4) in favour of the alternative hypothesis i.e. that a significant difference was evident between the raters' diagnosis of the child's difficulty following a three week interval. The Likelihood Ratio Chi Square was computed as the sample size was relatively modest. This statistic of 62.22 with 9 degrees of freedom, yielded a probability level of 0.0001. These figures confirm the previous Chi Square statistic. Fisher's Exact Test was computed and yielded a probability level of 4.36E-10. Cohen's Kappa was then computed. The value was 0.260 and yielded a probability level of 2.20e-006. The above computations are consistent with each other, and indicate significant results at all levels. It is, therefore, possible to reject the null hypothesis in favour of the alternative hypothesis (4).

5.6 Inter-Rater Agreement Of The Diagnosis Of The Child' Difficulty (2)

Finally the analyses of each of the raters was compared by means of a contingency table. If the raters had been in complete agreement, counts greater than zero would only be expected on the main diagonal.

5.6.1 Contingency table

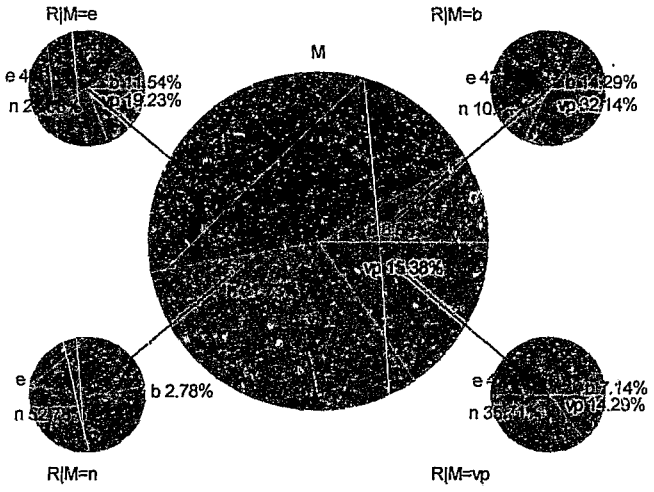
Table 15 Contingency table -comparison of analysis by rater

	Frequency Percent Row Pct Col Pct	R				
		both	emotional	neither	visual-perceptual	Total
M	both	4* 3.85% 44.44% 14.29%	3 2.88% 33.33% 11.54%	1 0.96% 11.11% 2.78%	1 0.96% 11.11% 7.14%	9 8.65%
	emotional	12 11.54% 26.09% 42.86%	12* 11.54% 26.09% 46.15%	16 15.38% 34.78% 44.44%	6 5.77% 13.04% 42.86%	46 44.23%
	neither	3 2.88% 9.09% 10.71%	6 5.77% 18.18% 23.08%	19* 18.27% 57.58% 52.78%	5 4.81% 15.15% 35.71&	33 31.73%
	visual-perceptual	9 8.65% 56.25% 32.14	5 4.81% 31.25% 19.23	0 0.00% 0.00% 0.00	2* 1.92% 12.50% 14.29	16 15.38%
	Total	28 26.92%	26 25.00%	36 34.62%	1 13.46%	104 100.00%

* Figures marked with an asterix fall on the main diagonal

5.6.2 Pie chart

Figure 5 Pie chart: Rater M vs Rater R



The central pie in this instance represents M's diagnoses and the peripheral pies represent Rs diagnoses for the corresponding cases, (i.e. of the 44.23 % of the sample that M analysed as emotional, R rated 46.15 % as falling into the emotional category and 23.08% as experiencing neither difficulty, 11.54% with both difficulty and 18.23% as experiencing visual-perceptual difficulties.

In order to establish the strength of agreement between raters concerning their diagnosis of the child's difficulty, a number of statistical techniques were applied. Firstly a contingency table (Table 15) was constructed which provides a graphic illustration of the raters' analyses. A sum of the counts along the main diagonal indicate that the raters agreed on the diagnosis of the child's difficulty in just 37 of the 104 cases (35 per cent), demonstrating a low level of agreement between the raters.

The breakdown of contingency table 15 is as follows: R diagnosed 26.92 percent of the protocols as falling into the BD category. This is in contrast to M's classification of only 8.65 percent. R classified 25 percent of the protocols into the ED category, in contrast to M's 44.23 percent. R placed 34.6 percent of the protocols into the ND difficulty, in contrast to M's 31.73 percent. Finally, R classified 13.46 percent of the protocols into the VPD category, in contrast to M's 15.38 percent. In order to make these percentages more meaningful, it is now necessary to consider the particular cases in which M and R agreed and in which cases they disagreed. This constitutes the core of the study.

Of the 26.92 percent of the protocols classified by R to the VPD category, M is in agreement with her in only 14.29 percent of the cases. M on the other hand considered that only 8.65 percent of the protocols exhibited BD. This indicates a tendency of R to classify far more cases than M as experiencing both difficulties. It also shows a tendency of M to use the "both" category fairly sparingly.

Of the group of 26.9 percent classified by R into the BD category, M classified 14.29 percent of this group into the same category. He placed 42.86 percent of this group into the ED category and 10.71 percent of this group into the ND category. He classified 32.14 percent of this group into the VPD category.

As mentioned, M classified only 8.65 percent of the total sample into the BD category; R was in agreement with M in 44.44 percent of this sample. She had classified 33.33 percent of M's sample into the ED category; 11.11 percent of this sample into the ND category, and 11.11 percent of M's sample into the VPD category.

R classified a total of 25 percent of the sample into the ED category. Of this group M was in agreement with her in 46.15 percent of the cases (just under half). M placed 11.54 percent of this sample into the BD category. He classified 23.08 percent of the sample into the ND category, and 19.23 percent of the sample into the VPD category. The fact that the raters appear to be in agreement in only 46.15 percent of the cases is considered by the researcher to indicate a very poor relationship.

M classified a total of 44.23 percent of the total sample into the ED category. Of these R is in agreement with him in 26.09 percent of the cases. She placed 26.09 percent of this sample into the BD category; 34.78 percent into the ND category; and 13.04 percent into the VPD category. R classified 34.62 percent of the total sample into the ND category. Of this group M was in agreement with her in 52.78 percent of the cases. However, he placed 2.78 percent of this sample into the BD category; and 44.44 percent into the ED category. M classified 31.73 percent of the protocols into the ND category. Of this group, R placed 57.58 percent into the same category; 15.15 percent into the VPD category; 18.18 percent into the ED category; and 9.09 percent into the BD category. R classified 13.46 percent of protocols as falling into the VPD category. Of these M considered that only 12.50 percent of this sample also fell into this category. He classified 42.86 percent of this group into the ED category, 15.15 percent into the ND category and 11.11 percent into the BD category.

M classified 15.38 percent of the total protocols into the VPD category. Of these, R was in agreement with him in only 14.29 percent of the cases. She considered 32.14 percent of this group as falling into the BD category; 19.23 percent into the ED category and no cases into the ND category.

In summary, as far as the BD category is concerned, M and R agreed in 3.85 percent of the cases. With regards to the ED category they agreed in 11.54 percent of the cases. The percentage of agreement rises to 18.27 percent as far as the ND category is concerned. As far as the VPD category is concerned, they are in agreement in only 12.50 percent of the cases. These figures illustrate that the strength of the relationship between the raters is minimal.

It would appear that a child who is assessed by the clinician M using only the Bender Gestalt and HFD is statistically most likely to receive a diagnosis of ED, while the child assessed by R is most likely to be classified with ND, i.e. with no diagnosis.

Following the construction of the contingency table, The Chi-Square, Fisher's Exact and Cohen's Kappa measures of association were computed. These are displayed below.

5.6.3 Statistics for inter-rater agreement

Table 16 Statistics for inter-rater agreement - Chi Square, Likelihood Ratio Chi Square, Fisher's Exact Test, Cohen's Kappa

Statistic	DF	Value	Prob
Chi-Square	9	23.307	0.006
Likelihood Ratio Chi-Square	9	28.499	0.001
Fisher's Exact Test (2-Tail)			1.06E-03
Cohen's Kappa		0.124183	
Observed Agreement		0.3557692	
Expected Agreement		0.2644231	
Normal Approximation		2.255014	0.01206621

The above statistics all have significance probabilities of less than 0.05. It is therefore possible to conclude that a significant difference was evident between the analysis arrived at by each of the raters' analyses. It is, therefore, necessary to reject the null hypothesis (2) in favour of the alternative hypothesis that there was a significant difference evident between the raters' diagnoses of the children's difficulties.

We may, therefore, conclude that even though there was a consistency between the raters in so far as their scoring of the HFD and Bender Gestalt tests were concerned, this broke down when they were required to make a "diagnosis" of a child's difficulty based on the two tests.

CHAPTER SIX: DISCUSSION

6.1 Introduction

The main aim of this research was to establish whether clinicians are usefully applying the Bender Gestalt and Human Figure Drawing in clinical practice in order to make appropriate referrals It was decided that one way of establishing this would be to determine whether two independent clinicians could achieve a sufficiently high level of agreement in terms of their scoring and diagnoses of the HFD and Bender protocols. The strength of the agreement between the raters was considered by the researcher to constitute an important aspect of the reliability of both the tools and of the raters.

Reliability was established in two ways : (a) by examining the strength of agreement between two clinicians in terms of their scoring and diagnoses (inter-rater reliability) and (b) by examining the consistency of each individual clinician's scores and diagnoses over two rating sessions (intra-rater reliability).

In order to create a statistically viable project, a number of null hypotheses were constructed to serve as the basis for the study. Each of these null hypotheses rested on the assumption that no significant difference would be found between the raters' scores and between their diagnoses. Further, it was postulated that there would be no significant difference between each individual rater's scores and their diagnoses from Test 1 to Test 2.

The results of the study, graphically displayed in the previous chapter, generated both significant and non significant results. On the one hand these results indicate a significant difference both between the two raters in scoring the Bender Gestalt Test and in diagnosing the children's difficulties; and on the other hand, a significant difference in diagnosis from Test 1 to Test 2. The statistical analyses were however unable to find a significant difference between the raters and from Test 1 to Test 2 as far as the HFD was concerned. These results are discussed in detail below. Following the discussion are comments on the limitations of the study as well as suggestions for future research.

6.2 Inter-rater reliability in the use of the Bender Gestalt Test

In the exploratory data analysis an ANOVA was conducted for the Bender Gestalt Test. This established that a significant difference was evident at the 5 percent level. (Table 6). It was then necessary to explore further the nature of the significance. An ANOVA was conducted which aimed to establish whether the apparent difference was a result of pre/post test effects within the raters themselves (intra-rater differences) or a result of inter-rater differences (pre/post test). In Table 5 it is evident that the only significant factor appears to be the rater effect (Pr>F 0.0174). This indicated that there was a significant factor between the two raters in terms of their scoring of the Bender Gestalt Test. It does not, however, specify the nature of this difference.

Given the fact that the differences occurred at the level of the rater with regards to the Bender Test, we compare the BMA means for each rater (Table 6), and find a significant difference at the level of mean difference (M's mean is -0.417 and R's mean is -0.985 indicating a mean difference of .57). This indicated that on average R assigned a lower score to the Bender protocols than did M. This difference was at the level of roughly half a year. The implication of a generally lower Bender age is that R regarded the children as experiencing greater visual perceptual difficulties than did M in comparison to their chronological age.

The reader is referred to Koppitz's inter-rater reliability studies which established correlation coefficients ranging from .79 - .9, with 81 percent exceeding .90. Thus a stronger relationship was evident in Koppitz's studies than in this research, where a correlation of .70 was noted. However it is also important to note Salvia and Ysseldyke's (1978) critique of Koppitz's studies. They observe that five of the nine reliability studies are on kindergarten children only, and only one of twenty five reported coefficients exceeds the standard of .90 recommended for tests that result in important decisions about children's functioning being made.

This has a number of implications for clinical practice. It could suggest that the differences are a result of what shall call be called "rater" differences. This refers to the possibility that the raters may not have applied the scoring manual in a sufficiently rigorous manner in order for them to arrive at the same scores. Each clinician is likely to enter an assessment situation bringing their own personal style to bear. One clinician will spend a great length of time scoring a protocol while another clinician will spend a relatively short time scoring a protocol. Each clinician will score the protocols with varying degrees of precision. Human error and inconsistencies are also likely to account for differences in scores. All the above are likely to affect the scores arrived at by the clinicians.

It is also useful to take into consideration the fact that the observed rater differences may be a reflection of different levels of expertise between the raters. It is therefore suggested that a more rigorous level of training be conducted at university level in order to standardise the scoring of this tool.

An analysis of the manual indicates that it demands a fairly high level of rigour. For example it requires that the clinician use a ruler and protractor when measuring angles or size. In some instances the clinician is required to count the number of dots in a series, or note whether angles do or do not intercept. This would appear to be straightforward. However, a certain degree of confusion may occur when the clinician refers to illustrations of scoring examples; as items that the manual indicates should be scored, may not actually feature in the example. The clinician is then left with a degree of uncertainty about whether to score the item or not. Thus the examples may in fact lend themselves to different interpretations.

It is further suggested that the language of the manual is insufficiently precise and leaves too much room for subjective clinical interpretation For example it states in page 15 of the scoring manual (Koppitz) that "since the Scoring System is designed for young children with as yet immature fine motor control, minor deviations are ignored". Here the terms "minor" and "deviation" are not further defined, leaving it entirely to the clinician to supply the content of these terms. This vagueness may serve to confuse rather than clarify. The researcher suggests that these factors may have been relevant in accounting for the seven month difference between M and R's scores.

6.3 Intra-rater reliability - scoring of the Bender Gestalt and HFD Tests

The reliability with regards to the scoring of the Bender Gestalt and HFD from Test 1 to Test 2 are considered together as consistent results were evident in both cases.

As indicated by the statistical studies in Chapter five there was a high level of intra-rater reliability from Test 1 and Test 2 as far as both the Bender Gestalt and HFD tests were concerned. It was therefore not possible to reject the null hypotheses (3a, 3b).

The relatively high level of agreement between Test 1 and Test 2 for the scoring of both the Bender Gestalt and HFD enabled the following conclusions to be drawn: A child who is referred for and receives an assessment on a certain date, and is administered the Bender Gestalt and/or HFD tests as one or two of the assessment instruments used, is likely to receive a relatively similar score for his or her test if the same clinician, at another date, tests the same child with the same tools.

It is noteworthy that the consistent results were obtained where the only information made available to the clinicians was gender and age. This suggests that the scoring of the protocols themselves were highly independent of any other knowledge of the child. It is not necessary to know, for example, that a child who is referred for assessment necessarily has a history of occupational therapy related difficulties.

This is relevant in the South African context where children may be brought for an assessment by a figure other than the biological parent/s or in many instances may be brought by a family friend or sibling, and no background history or description of the nature of the child's difficulty is necessarily available. This would indicate a validation of the use of these tools in the South African context as possible screening devices for emotional and visual-perceptual difficulties. However an important observation needs to be made here [illegible] ers achieved a fairly high level of consistency from Tests 1 to 2 in ter [illegible] of the protocols for both the Bender Gestalt and HFD, their diagnoses [illegible] ureliable (Tables 13, 14, 15, 16). The important point being made here is tha [illegible] raters were instructed to base their diagnoses on the results of the Bender Gestalt and HFD protocols. These discrepancies are explored further in the next section.

6.4 Inter-rater agreement - in the use of the Human Figure Drawing

As with the Bender test, the analysis of inter-rater agreement was achieved by comparing the scores of each rater to the other. The initial statistical analysis for the scoring of the human figure drawing finds no significant difference between the raters scoring or between Tests 1 and 2. The statistical model was not significant, indicating insufficient evidence to show a significant difference between the raters (Pr>0.4927). This means it was not possible to reject the null hypothesis (1b) in this instance. Summary statistics (Table 10) indicated that there was no significant mean difference between the raters. However M displays a smaller standard deviation suggesting less variance in his scoring and, therefore, more consistency within himself than the R. The trend for rater M to be more consistent than rater R was evident in the scoring of both tests.

Pearson's Correlation Coeffiicent shows a correlation of 0.64 at the 0.0001 level of significance (Table 11). This indicates a relatively high level of agreement between the raters regarding the scoring of the HFD. This positive relationship indicates that the two raters identified the same protocols as exhibiting emotional indicators.

The reader is referred to Strumpfer and Nicholas's (1963) study where inter-rater correlation's for six different aspects of figure drawings (overall quality, adjustment, sexual differentiation, maturity and body image disturbances) ranging from .79 - .97 were observed. This correlation is higher than that found in this research and may have resulted from the fact that only six criteria (as opposed to thirty) were employed. However, these criteria appear to be more encompassing and perhaps more difficult to rate than Koppitz's thirty criteria.

Research by Cassel et al (1958) found a high level of inter-rater correlation (0.9) regarding the presence or absence of emotional signs on human figure drawing protocols. after three training sessions in the use of the tool had been conducted. Although Cassel's research did not apply the Koppitz criteria, the raters in his study, and in this research, were nevertheless requested to make similar judgements, i.e. they were asked to note whether a certain emotional indicator was present or absent.

Cassel's correlation of .90 is higher than that achieved in this study. This could be explained by the fact that Cassel's raters were provided with a number of training sessions in the use of the tool, which may have resulted in a standardisation of their techniques. Had the raters in this research been fully trained in the use of the Koppitz criteria prior to the conducting of the rating sessions themselves, a higher inter-rater correlation may have been observed.

6.5 Intra-rater reliability of the diagnosis of the child's difficulty

The statistical results indicated that while the raters were able to arrive at a reasonable level of consistency with regards to their scoring of the protocols, this was not the case as far as their diagnoses of the child's difficulty was concerned. The reader is referred to Tables 13 and 14 where is was noted that the null hypothesis (4) that no significant difference would be evident between the raters' diagnosis of the child's difficulty following a three week interval, was rejected in favour of the alternative hypothesis in this instance. This was that a significant difference would be evident. This lack of agreement was graphically represented by the contingency table (Table 13).

As illustrated by the contingency table, the BD category was more extensively used at Test 2 than at Test 1. This sample was largely derived from protocols that had been assigned to the ED category at the first session. Hypothetically, a child who was diagnosed with, for example, a VPD difficulty at his/her first assessment (and therefore, may have been referred for occupational therapy), had a fairly high chance of being rediagnosed -at a repeated assessment - as experiencing both a visual-perceptual and emotional difficulty. Similarly a child classified with an emotional problem at his/her first assessment has a high chance of being re-diagnosed with both a visual-perceptual and emotional difficulty at a second assessment. This could possibly result in referrals to both occupational and play therapy.

This researcher considers that only children from affluent families are likely to be enrolled in both therapies since they are costly. The vast majority of South African children are unlikely to have the financial means to be able to receive both play and occupational therapy. Parents may, therefore, decide to send their child to either a play therapist or to an occupational therapist (not necessarily both).

This may result in neither difficulty being effectively addressed. Alternatively, if a child is enrolled in both therapies, he/she may feel overwhelmed and may begin to show signs of pathologising him/herself. A third scenario is that parents may not be able to afford therapy of any kind, and feel helpless with regards to their child's difficulty. Or, possibly, the parent may choose the cheaper therapy (usually OT) and an underlying emotional difficulty that inhibits effective learning may not be addressed.

It is noticeable that a percentage of the protocols that were classified into the ND category at Test 1 were re-classified into the ED category at Test 2. This indicates a tendency of the raters to diagnose an emotional difficulty where none was previously noted. This could suggest that a genuine difficulty that was not picked up at Test 1 was identified at Test 2. Conversely, it might indicate that an emotional difficulty diagnosed at Test 2 could have been incorrect. The possible consequences of this are: an unnecessary referral, the incurement of costly play or group therapy, feelings of uncertainty and helplessness on behalf of both the parents and the child, all on the basis of a dubious diagnosis.

It is also noticeable that although the total percentage of children classified into the ND category shrank from 40.38 percent of the total sample to 25.96 of the total sample at Test 2, it is evident that a large percentage of the children (81.48 percent) classified into the ND category were, in fact, the same group of children. Therefore, although the ND category shank, it is evident that the highest level of agreement occurred in this category, perhaps indicating that the ND category was the most reliable as far as the clinicians ratings were concerned.

An equal percentage of the protocols which had been placed into the BD, ND and ED categories at the first session were re-classified by the raters into the VPD category at Test 2. This indicates a fairly radical re-categorisation as far as the raters were concerned, especially the shift from an emotional to a visual-perceptual difficulty. One wonders how, in a professional clinical setting, a child could be reclassified from one session to the next, and would, therefore, be referred for a vastly different type of therapy. For example a child with an emotional difficulty, who might have been referred for play therapy at the first session, was reclassified at the second session as experiencing a visual-perceptual problem and may be referred for occupational therapy.

If one extrapolates to clinical practice in general it would indicate that clinical diagnoses and referrals made on the ba is of these tools are often unreliable. It also indicates that these tools cannot be used as a basis for making diagnostic decisions, or that the tools cannot be utilised on their own, but must be used in conjunction with a number of other tools. It is also suggested that if clinicians are to continue to apply the Bender Gestalt and HFD, then they require far more rigorous training in the use of the tools.

While, overall, the VPD category appears to have expanded slightly at Test 2, nonetheless less than 50 percent of the actual children who were thus classified at Test 1 were classified into the VPD category at Test 2. Again, this illustrates the lack of concordance between the raters' diagnoses between the first and second rating sessions. At Test 1 a significant proportion of the protocols (40.38 percent of the total sample) were classified into the ND category. This suggests an initially conservative approach by the raters at Test 1. At Test 2 however, they tended to rate more protocols as falling into the ED category (33.65 percent of the total sample) When discussing these percentages it is important to bear in mind, however, that they refer to percentages, and not to specific children. The statistical figures and tabulations do not reveal the identity of the children. Therefore it is not possible to say that child "X" was rated with a visual perceptual difficulty at one assessment and an emotional difficulty at another. It is only possible from these statistics to comment on a general trend of the raters to be less discriminating as well as less conservative at Test 2 in comparison to Test 1. This observation is most noticeable as far as the BD category is concerned. At Test 1, 10.58 percent of the total sample were classified into the BD category. At Test 2, this grew to 25 percent of the total sample. This suggests a tendency of the raters to assign children to the BD category, rather than making what could be viewed as a more discerning diagnosis. This also indicates a tendency to pathologise the children at Test 2. It is difficult to determine why the shift in classification occurred, however, it is noteworthy that the shift was not made on the basis of the Bender Gestalt and HFD scores as they were consistent over time. Perhaps when in doubt, the raters decided to assign the protocols to the BD category in order to cover all possible bases.

In both Test 1 and Test 2, the VPD category appears to be the most under-utilised, indicating a tendency of the raters to be least likely to diagnose a child as experiencing a visual perceptual difficulty.

In summary, although the clinicians were able to achieve a fairly high level of consistency at the level of scoring between their first and second rating sessions, they were not able to achieve consistency in terms of their diagnoses.

The following deduction can be made: In terms making their diagnoses, the clinicians did not rely on the test scores per se. Had they done so this would presumably have been reflected in the statistical analyses i.e. a similar pattern would have been evident for both the scoring and the diagnoses sections. The fact that there was high agreement at the level of scoring but low agreement at the level of diagnosis must suggest that diagnoses were not made on the basis of score results. For example a low Bender score would be most likely to result in a diagnosis of a visual-perceptual difficulty, while a high HFD score would be most likely to result in placement in the emotional category. An age appropriate Bender score coupled with a low HFD score may have resulted in the child being classified into the ND category. Similarly, a low Bender score coupled with a high HFD may have resulted in the utilisation of the BD category.

It is troubling that the same clinician cannot arrive at the same conclusion for the same child at a repeated session, and we can therefore extrapolate that the same could be the case in clinical practice generally. The analogy of the doctor is appropriate here. For example, if I consult a skin specialist about a skin mole and she concludes that it is normal in all respects, I would most likely not pursue the matter any further. However, if I consult the same doctor regarding another matter three weeks later, and at this consultation she now diagnoses the mole as malignant, I would consider this to be very poor practice on behalf of the doctor. Just as we expect the same doctor to make the same diagnosis, we also expect different doctors (with similar professional training and expertise) to make the same diagnosis. As we expect the same psychologist to make the same diagnosis at different assessments, we except different psychologists with similar professional training and expertise to arrive at the same conclusions concerning the same child.

The tools may not by themselves provide sufficient or accurate material upon which to make useful diagnoses. We, therefore, need to question for what purposes clinicians are generally using these tools and also why they are almost routinely included in a test battery.

6.6 Inter-rater agreement in diagnosis

In order to establish the strength of agreement between the diagnosis of the raters (inter-rater agreement), the same statistical techniques were applied to the data as were used in establishing intra-rater reliability.

It is apparent that a child assessed by M using only the Bender Gestalt and HFD was most likely to receive a diagnosis of an emotional difficulty, while the child assessed by R was most likely to be classified as experiencing neither difficulty, i.e. with no diagnosis.

As discussed, a high inter and intra-rater agreement was evident in terms of the scoring of the HFD and a high intra-rater agreement was evident in the scoring of the Bender Gestalt. This is considered by the researcher to be fairly unremarkable. The lack of inter-rater agreement in terms of the scoring of the Bender Gestalt and in terms of the inter and intra-raters' clinical formulations is, however, troubling. While a certain amount of human inconsistency was to be expected, this study has revealed the inconsistency between the raters' clinical formulations was so large as to raise serious doubts as to the practice of clinical assessment in this form.

In order to practice as a psychologist in South Africa, it is necessary to undertake at least six years' professional and practical training. It would be expected that given this level of professional training and clinical experience, psychologists would be able to score an assessment tool with a reasonable level of consistency. As we have seen, this was largely the case. However, the fact that they achieved fairly similar *scores* in most instances, is insignificant, unless these scores are placed within a meaningful context. For example, merely stating that a child who is seven years of age has a developmental age equivalence of 6.0 to 6.5 years in terms of the Bender Test, has no significance in and of itself. Similarly, stating that a child's HFD exhibits two emotional indicators means little on its own.

It is only when one makes the step from scoring a test to making a comment on what that score means for the child's development or level of functioning, that the score becomes clinically significant. For example stating that two emotional indicators revealed by the HFD suggest a shy or aggressive child, one is attaching meaning to numbers.

It is a fairly simple task to observe whether teeth are present in a drawing or not, or whether a figure shows a rotation of 45 degrees or not, but to extrapolate from these markers to a useful clinical formulation of the nature of the child's difficulty, requires a high level of expertise. To reiterate, computing a score is one activity. Extrapolating, nuancing and diagnosing is a far more complex task.

It is also notable that the two raters tended to favour different diagnostic categories. M showed a bias towards utilising the ED category, while R was most likely to utilise the ND category. These discrepancies could have sprung from the fact that M and R may have been exposed to different theoretical schools of psychology. It is therefore possible that in addition to differing levels of clinical expertise accounting for the apparent discrepancies between the raters, a factor that also needs to be considered is the clinician's prior training and theoretical grounding.

As discussed in the literature review, a vociferous body of criticism has been levelled at the use of projective tools, in particular the DAP and HFD. This has been the case not only because of seemingly unsubstantiated conclusions that are drawn from HFD's, but also because of a strong subjective component that seems to accompany their use by some clinicians. This research bears out the fact that the when tools such as the HFD are used to make clinical judgements of children's difficulties, different clinicians are not able to make these judgements with a sufficiently high level of reliability.

The research shows the mean age of this sample of children to have been approximately nine years of age, i.e. grade three level. This is in line with the researcher's experience in various clinical settings where a child is likely to be referred for an assessment at roughly this age. This may be the case because it is frequently only in grade three that certain learning difficulties become apparent. If the child's problem is not effectively identified and remediated at this pivotal stage of his/her development, the child's future school, career and indeed life opportunities could be adversely affected.

This research has indicated that two clinicians, who achieved fairly consistent scores, achieved very inconsistent diagnoses. By extension, it is possible to suggest that a similar pattern of inconsistency would be evident amongst clinicians in practice who use these tools to make similar judgements. If the results of assessments are not sufficiently reliable, it begs the question as to whether the time and considerable expense involved in clinical assessment (and potential damage to the child through misdiagnosis), is warranted.

6.7 Limitations of the study and suggestions for future possible research

This research report has largely attributed the differences in scoring and clinical formulations to the clinicians themselves. By extension it has questioned clinicians in practice who use these tools as the basis for making similar clinical formulations. However, locating the problem solely with the raters has in part sprung from the design of the study which was weighted towards finding fault with the raters.

The study made the following assumptions: A low Bender Score indicated the presence of occupational therapy difficulties; a high HFD score indicated the presence of emotional difficulties; a low Bender score coupled with a high HFD score indicated the presence of both difficulties, and an age appropriate Bender score coupled with a low HFD score indicated the absence of either difficulty. As discussed, the raters were requested to assign the children's protocols to one of four categories. There are a number of problems with these assumptions which are discussed below.

As explored in the review of the literature, controversy surrounds the use of these tools. Concerning the Bender Gestalt, the reader is referred to the study of Shapiro and Simpson (1995) which found no significant evidence of a relationship between Bender Gestalt test errors and specific academic skills, i.e. no evidence that Bender Gestalt errors warrant a referral to an occupational therapist to remediate the visual-perceptual difficulties. Other research by Fuller et al (1998), suggest that there is little evidence to support the construct validity of the Bender Gestalt. Buckley (1978) also refers to studies that find no conclusive evidence to support the use of the Bender Gestalt in predicting school achievement or identifying neurological or *visual-perceptual impairment.* It is therefore quite possible that requesting the raters to diagnose a visual-perceptual difficulty on the basis of a low Bender score rested upon a controversial and unproven correlation.

In a similar way, research into the validity and reliability of the HFD Koppitz criteria is also equivocal. Research by Tharinger and Stark (1990) (referred to in the literature review) found that the thirty Koppitz emotional indicators were not able to distinguish between emotionally disturbed children and non-disturbed controls. Therefore, requesting the raters to diagnose an emotionally based difficulty on the basis of a high HFD score (possibly resulting in a referral to play therapy or similar) may also have rested upon an unproven assertion.

The disparity between the raters might therefore have been caused by the lack of a direct relationship between the test scores and the diagnostic categories. However, questioning the validity of the diagnostic categories was not the primary aim of this report. Future research is urgently needed to explore the validity of the Bender Gestalt as a test of visual-perceptual functioning, and the HFD as a measure of emotional functioning.

A further inadequacy of the diagnostic categories as they were constructed in this research, is that they did not take into account the complexity of the relationship between visual-perceptual and emotional functioning. For example, an emotional difficulty may present as a difficulty in the visual-perceptual domain. Similarly, a visual-perceptual deficit may present as a socio-emotional difficulty. It is further possible that there is in fact a causal relationship between visual-perceptual and emotional functioning. An underlying emotional difficulty may in part be the cause of a visual-perceptual difficulty and vice versa. In clinical practice neat categorisations into VPD, ED etc. are probably rare. The situation is often more nuanced and decisions more tentative.

A criticism of the diagnostic categories as they stand, therefore, is that they could have been somewhat artificial and arbitrary, and may have constrained the clinicians' judgements. A future study might allow testers the latitude to use their own diagnostic categories as opposed proscribing the possibly flawed ones used in this research. In such a future study the clinicians would be asked to state what information they gleaned from the Bender Gestalt and/or HFD tests, and what sort of judgements they feel confident of making on the basis of these tests. The clinical formulations and judgements of the different clinicians would then be compared.

Earlier in the discussion, it was observed that the rater M appeared to be more discerning and precise in his diagnoses and scoring than the rater R. This study did not however establish whether either rater was in fact accurate in his/her diagnoses. In other words, it was not possible to conclude whether for example M's classification of a child's diagnostic protocol to the visual-perceptual category was in fact an accurate classification. The research might have in fact merely have compared one set of misclassified protocols to another set. Future research might employ more raters: This would enable statistical trends to be established. For example, if multiple raters were used and 80 percent of them classified a group of the sample to the ED category etc., this could provide far greater confidence in the reliability of the raters as well as greater validity to the diagnostic categories.

Other limitations of the research included its relatively small sample size, and the fact only two rating sessions were held. Increasing the sample size, the number of rating sessions, and the time interval between them would all lend statistical weight to this type of research and would establish the level of inter and intra-rater agreement more convincingly.

Although this study only examined each test as used for its primary function, i.e. Bender Gestalt for determining visual-perceptual and the HFD for determining emotional functioning, it is noted in the literature review that both tests also have secondary functions. Lauretta Bender included various criteria of emotional functioning for use with the Bender Gestalt Test and various researchers use information from the HFD to establish visual-perceptual and intellectual functioning (Goodenough, 1926). A future study could compare the information gained from a Bender Gestalt protocol to the information gained from an HFD protocol concerning visual-perceptual functioning only. This would enable the researcher to establish the more reliable and useful tool in so far as visual-perceptual functioning is concerned, and establish the nature of the correlation between these tools applied in this way. Similariy, the study could compare information gained from an HFD protocol to that gained from a Bender Gestalt protocol in so far as emotional functioning only is concerned., and establish the nature of the correlation between the tools in this regard. Again this would enable the research to establish which tool provides more useful and reliable information.

The complex relationship between visual-perceptual and emotional functioning is illustrated in the following example. A child for example may have drawn a very small, unbalanced, human figure, with poor integration of body parts (possible indicators of poor visual-perceptual functioning). However, in this study the child's HFD protocol would be scored for emotional indicators only. He/she may nevertheless have been able to adequately reproduce the Bender Gestalt figures, and achieve an age appropriate score. Therefore, his/her possible visual-perceptual difficulties (as reflected by the HFD) may be masked by the Bender Gestalt test. This research could perhaps have been considerably enhanced therefore had the clinicians been requested to gauge visual-perceptual and emotional functioning from both tools.

It is also possible that these tools could be more effectively employed if clinicians were provided with considerably more training in the possible implications of low and high test scores. Considerations of university training indicates that very little time is devoted not only to their application, but also to a linkage of their theoretical underpinnings and clinical applications. If clinicians received more comprehensive training it would be worth repeating a study of this nature to explore whether reliability would be higher and would give more weight to either their continued use, or their abandonment.

6.8 Conclusion

In a country where many children have been subject to an historically inferior education system, and are presenting with learning difficulties, appropriate problem identification and remediation are vital. This is all the more so considering the socio-economic backdrop, where the vast majority of South African families will struggle financially to afford the therapies that clinicians suggest. Inaccurate identification of a child's difficulty could lead to inappropriate referrals, the squandering of precious resources, and in a worse case scenario, to the child being adversely affected.

A lack of consistency between the clinical formulations of psychologists might imply that children's difficulties are not being accurately identified. This research addressed that concern. It aimed to establish the inter-rater reliability of two independent clinicians' scoring of the Bender Gestalt Test and the Human Figure Drawing, and the reliability of their diagnoses based on the results of these two tools. This research also aimed to establish the intra-rater reliability of their scoring and diagnoses at a second scoring session held three weeks after the first. The Bender Gestalt and HFD were selected as they are widely used and easy to administer tools that appear to be almost always included in assessment batteries in South Africa.

This research found a low level of agreement between the raters in terms of their diagnoses, and their scoring of the Bender Gestalt Test. As far as the diagnoses were concerned, low intra-rater reliability was also observed. However this research established reasonable inter and intra-rater reliability for the scoring of the HFD, and reasonable intra-rater reliability for the scoring of the Bender Gestalt.

These results indicate that there is cause for concern regarding the reliability of clinicians interpretations based on Bender Gestalt and Human Figure Drawings. Considering that both these tools are widely used in South Africa, further research to establish their validity is warranted. In addition a more comprehensive reliability study should be conducted. If significant results are established, more rigorous training in the clinical use of these tools is urgently indicated.

REFERENCE LIST

Anastasi, A. (1982) Psychological Testing. Fifth Edition. New York: Macmillan

Bannantyne, A. (1971) Language, Reading and Learning Disabilities. Springfield, H: Charles C Thomas

Bender, L. (1938) A Visual Motor Gestalt Test And Its Clinical Use. New York: The American Orthopsychiatric Association.

Billingslea, F. (1948) The Bender Gestalt Test: An objective scoring method and validating data. Journal of Clinical Psychology, Vol. 5

Buckley, P.D. (1978) The Bender-Gestalt Test: A review of reported research with school age subjects. Psychology in the Schools, 3, 327-338

Cassel, R. , Johnson, H. & Burns, W. H. (1958) Examiner, ego defence, and the H-T-P Test. Journal of Clinical Psychology, Vol 14

Chittooran, M. & Miller, T. (1998) Best Practices in Assessment of Children. Psychological Assessment of Children: Best Practices for School and Clinical Settings. Second Edition, Vance. B. (Ed) New York: John Wiley and Sons Inc.

Cox, M. (1993) Children's Drawings of the Human Figure. East Sussex: Lawrence Erlbaum Associated Ltd.

Drever, J. (1979) The Penguin Dictionary of Psychology. Middlesex: Penguin.

Fleiss, J.L. (1981) Statistical Methods For Rates And Proportions London: J Wiley and Sons

Fuller, G.B, & Vance, B. (1995) Interscorer reliability of the modified version of the Bender-Gestalt Test for Preschool and Primary School Children. Psychology in the Schools Vol 32

Fuller, G., Awadh, A., & Vance, B. (1998) Assessing perceptual motor skills Psychological Assessment of Children: Best Practices for School and Clinical Settings. Second Edition, Vance. B. (Ed) New York: John Wiley and Sons Inc.

Gabel, S., Gerald, O. & Butnik, S. (1986) Understanding Psychological Testing In Children. A guide for health professionals. New York: Plenum Publishing.

Goldstein, D. & Britt, T. (1994) Visual Motor Coordination and Intelligence as Predictors of Reading, Mathematics, and Written Language Ability. Perceptual and Motor Skills, Vol 78

Goodenough, F. L. (1926) Measurement of Intelligence by Drawings. New York: Harcourt Brace and World

Hallahan, D. & Cruickshank, W. (1973) Psychoeducational foundations of learning disabilities. Englewood Cliffs, NJ: Prentice Hall

Hammer, E. (1968) The Clinical Applications of Projective Drawings. Springfield: Charles C. Thomas

Harris, D. B. (1963) Children's Drawings as Measures of Intellectual Maturity. New York: Harcourt Brace and World

Howell, D. (1989) Fundamental Statistics For The Behavioural Sciences. Boston: PWS - Kent Publishing Company

Kahill, S. (1984) Human figure drawing in adults: An update of the empirical evidence Canadian Psychology: 25, 269-292

Kamphaus, R. W. & Pleiss, K.L. (1991) Draw-a-person technique: Test in search of a construct. Journal of School Psychology. Vol 29

Kaufman, A. S. (1979) Intelligence Testing With The WISC-R. New York: John Wiley

Kline, P. (1993) Personality: The Psychometric View. London: Routledge

Knoff, H. M., Cotter, V., & Coyle, W. (1986) Differential effectiveness of receptive language and visual-motor assessments in identifying academically gifted elementary school student. Perceptual and Motor Skills. Vol 63

Koffka, K. (1935) Principles of Gestalt Psychology. London: Kegan Paul

Koppitz, E. (1968) Psychological Evaluations of Children's Human Figure drawings. New York: Grune & Stratton

Machover, K. (1949) Personality Projection in the Drawing of the Human Figure. Springfield: Charles C. Thomas

McKay, M.F. & Neale, M.D. (1985) Predicting early school performance in reading and handwriting using major "error" categories from the Bender-Gestalt test for young children. Perceptual and Motor Skills, Vol 60

Mcneish, T.J. & Naglieri, J.A. 1993 Identification of individual with serious emotional disturbance using the Draw A Person: Screening procedure for emotional disturbance. The Journal of Special Education. Vol. 27/No.1 pp. 115-121.

Mortensen, K. V. (1991) Form and Content in Children's Human Figure Drawings. New York: New York University Press

Naglieri, J.A. (1988) Draw A Person: A Qualitative Scoring System. San Antonio, TX: Psychological Corporation

Nielson, S. & Sapp, G.L. (1991) Bender-Gestalt developmental scores: Predicting reading and mathematics achievement. Psychological Reports Vol 69

Pascal, G. & Suttell, B. (1951) The Bender Gestalt Test: Quantification and Validity for Adults. New York Grune & Stratton

Porteous, M. (1996) The Use of the Emotional Indicator Scores on the Goodenough-Harris Draw-a-Person Test and the Bender Motor-Gestalt Test to Screen Primary School Children for Possible Emotional Maladjustment. European Journal of Psychological Assessment, Vol 12. 1, pp. 23-26

Rabin, A. (1986) Projective Techniques For Children. New York: Macmillan

Reichenberg, N. & Raphael, A. (1992) Advanced Psychodiagnostic Interpretation of the Bender Gestalt Test. New York: Praeger

Richter, L., Griesel, R.D., & Wortley, M. (1989) The Draw a Man Test. A 50th year perspective on drawings done by black South African children. South African Journal of Psychology. Vol 19 (1)

Roback, H. (1968) Human Figure Drawings: Their utility in the clinical psychologist's armamentarium for personality assessment Psychological Bulletin, Vol 70, No. 1

Rudenberg, S. L., Jansen, P. & Fridjon, P. (1998) The effect of exposure during an ongoing climate of violence on children's self-perceptions, as reflected in drawings. South African Journal of Psychology Vol. 28, No. 2

Sabatino, D., Fuller, C, & Altizer, E. (1998) Assessing current psychological status. Psychological Assessment of Children: Best Practices for School and Clinical Settings. Second Edition, Vance. B. (Ed) New York: John Wiley and Sons Inc.

Salvia, J. & Ysseldyke, J. (1988) Assessment in Special and Remedial Education. Fourth Edition. Boston: Houghton Mifflin Company

Savage, R.D. (1968) Psychometric Assessment of the Individual Child. Harmondsworth: Penguin.

Shapiro, S, & Simpson, R. (1995) Koppitz scoring system as a measure of Bender-Gestalt performance in behaviourally and emotionally disturbed adolescents. Journal of Clinical Psychology, Vol 51, No 1

Strumpfer, D. J. W. & Nicholas, R.C. (1962) A study of some communicable measures for the evaluation of Human Figure Drawings. Journal of Projective Techniques. Vol 26

Stuart, A. & Ord, K. (1973) Kendall's Advanced Theory of Statistics - Vol. 2 – Classical Inference and Relationship. London: Edward Arnold

Swenson, C. (1957) Empirical Evaluations of Human Figure Drawings Psychological Bulletin, Vol. 70. No. 1

Swenson, C (1968) Sexual differentiation on the Draw-a-Person Test. Journal of Clinical Psychology, 11 37-40

Tharinger, K.J. & Stark, K. (1990) A qualitative versus quantitative approach to evaluating the Draw-A-Person and Kinetic Family Drawings: A study of mood and anxiety disorder children. Psychological Assessment: A Journal of Consulting and Clinical Psychology. Vol 2. 365-375

Tolor, A. & Brannigan, G. (1962) Research and Clinical Applications of the Bender-Gestalt Test. Springfield. Charles C. Thomas

Tolor, A. & Schulberg, H. (1962) An Evaluation of the Bender Gestalt Test. Illinois: Charles C Thomas Publisher.

Wilson, S & Reschly, D. (1996) Assessment in School Psychology Training and Practice. School Psychology Review, Vol 25, No. 1

Yama, M. (1990) The usefulness of human figure drawings as an index of overall adjustment. Journal of Personality Assessment Vol 54. (1)

Appendix A

SCORING SHEET – BENDER GESTALT

Figure	Error		Scored	BI Indicator	
				Significance Level 0.05	0.01
A	1 a	Distortion			
	1 b	Disproportion			
	2	Rotation			
	3	Integration			
1	4	Circles for dots			
	5	Rotation			
	6	Perseveration			
2	7	Rotation			
	8	Integration			
	9	Perseveration			
3	10	Circles for dots			
	11	Rotation	4		
	12 a	Shape lost			
	12 b	Line for dots			
4	13	Rotation			
	14	Integration			
5	15	Circles for dots			
	16	Rotation			
	17 a	Shape lost			
	17 b	Line for dots			
6	18 a	Angles for curves			
	18 b	Straight line			
	19	Integration			
	20	Perseveration			
7	21 a	Disproportion			
	21 b	Distortion			
	22	Rotation			
	23	Integration			
8	24	Distortion			
	25	Rotation			
		TOTAL			

4

BENDER NORMS

NORMATIVE DATA

Normative data for the *Developmental Bender Scoring System for Children* were derived from 1104 public school children representing 46 entire classes in twelve different schools located in rural, small town, suburban, and urban areas of the Midwest and Eastern States. Only those children were excluded from the normative population whose age was below 5 years or above 10 years and 11 months. The 46 classes included 10 kindergarten rooms, 13 first grades, 11 second grades, 5 each of the third and fourth grades, and two fifth grades. The distribution of the normative population by age and sex is shown on *Table 4*. The Bender Gestalt Test was administered individually to each child in

Table 4. Distribution of Normative Population by Age and Sex

Age	Boys	Girls	Total
5-0 to 5-5	37	44	81
5-6 to 5-11	81	47	128
6-0 to 6-5	89	66	155
6-6 to 6-11	103	77	180
7-0 to 7-5	95	61	156
7-6 to 7-11	62	48	110
8-0 to 8-5	39	23	62
8-6 to 8-11	37	23	60
9-0 to 9-5	32	33	65
9-6 to 9-11	28	21	49
10-0 to 10-5	15	12	27
10-6 to 10-11	19	12	31
Total	637	467	1104

school by a qualified psychologist; all Bender protocols were scored by the author according to the Developmental Bender Scoring System.

Mean Scores of Normative Population

The means of the Bender composite scores for the boys and the girls and for all children at each age level are presented on *Table 5*. The same data are shown on the graph on *Figure 1*. It can be seen that the Bender mean scores

Table 5. Bender Mean Scores by Age and Sex for Normative Population

Age	Boys	Girls	All Subjects
5	14.3	13.0	13.6
5½	10.0	9.3	9.8
6	8.3	8.0	8.1
6½	6.2	6.6	6.4
7	5.3	4.2	4.8
7½	4.9	4.4	4.7
8	3.9	3.6	3.7
8½	2.6	2.4	2.5
9	1.5	1.8	1.7
9½	1.6	1.5	1.6
10	1.5	1.7	1.6
10½	1.4	1.5	1.5

Appendix B

Scoring sheet for the Human Figure Drawing (adapted from Elisabeth Munsterberg Koppitz)

Quality Signs	**Scored**
1. Poor integration of parts of figure	_____
2. Shading of face	_____
3. Shading of body and/or limbs	_____
4. Shading of hands and/or neck	_____
5. Gross asymmetry of limbs	_____
6. Slanting of figure, axis of figure tilted by 15 degrees or more	_____
7. Tiny figure - 5.08 cm or less	_____
8. Big figures - 22.8 cm or more in height	_____
9. Transparencies	_____
(Subtotal)	_____

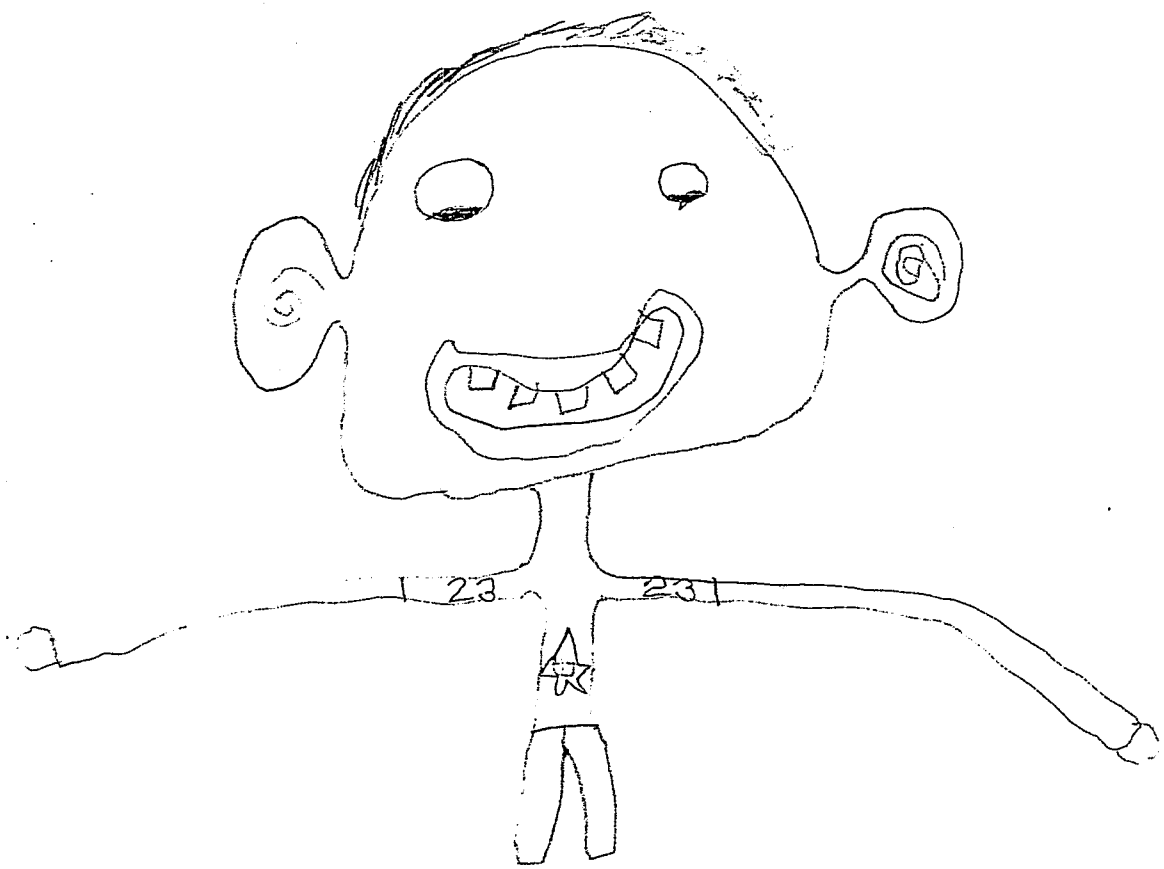
Special Features	
10. Tiny head, less than 1/10th of total figure in height	_____
2. Crossed eyes, both eyes turned in or out	_____
12. Teeth	_____
13. Short arms, arms not long enough to reach waistline	_____
14. Long arms, arms long enough to reach knee line	_____
15. Arms clinging to side of body	_____
16. Big hands, hands as large as face of figure	_____
17. Hands cut off, arms without hands or fingers	_____
18 Legs pressed together	_____
19. Genitals	_____
20. Monster or grotesque figure	_____
21. Three or more figures spontaneously drawn	_____
22. Clouds, rain, snow	_____
(Subtotal)	_____

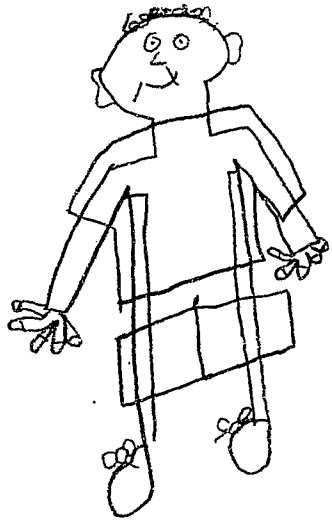
Omissions	
23. No eyes	_____
24. No nose	_____
25. No mouth	_____
26. No body	_____
27. No arms	_____
28. No legs	_____
29. No feet	_____
30. No neck	_____
(Subtotal)	_____

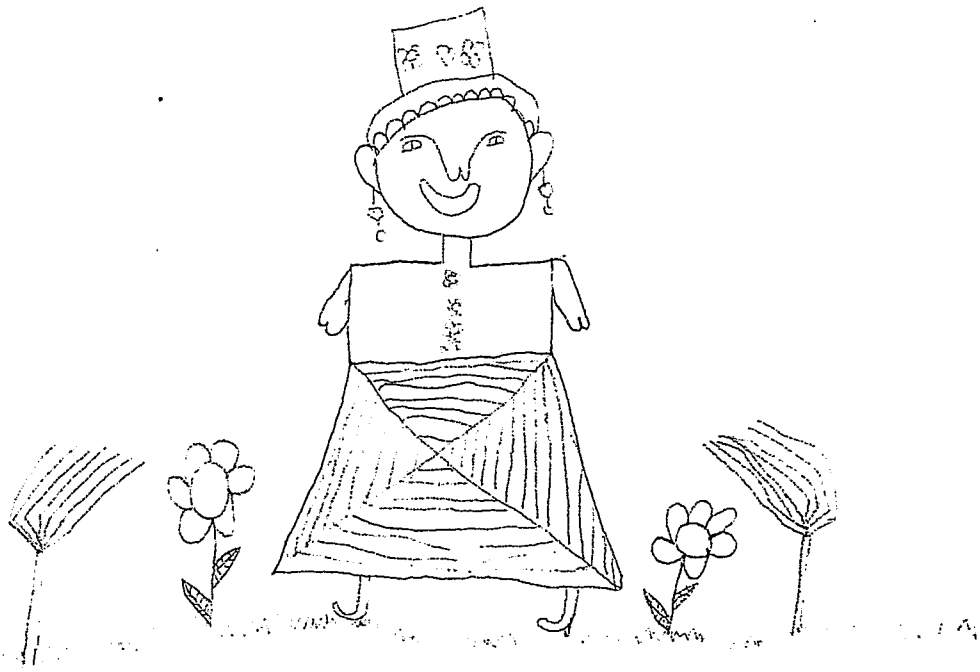
Total number of emotional indicators **(Total)** _____

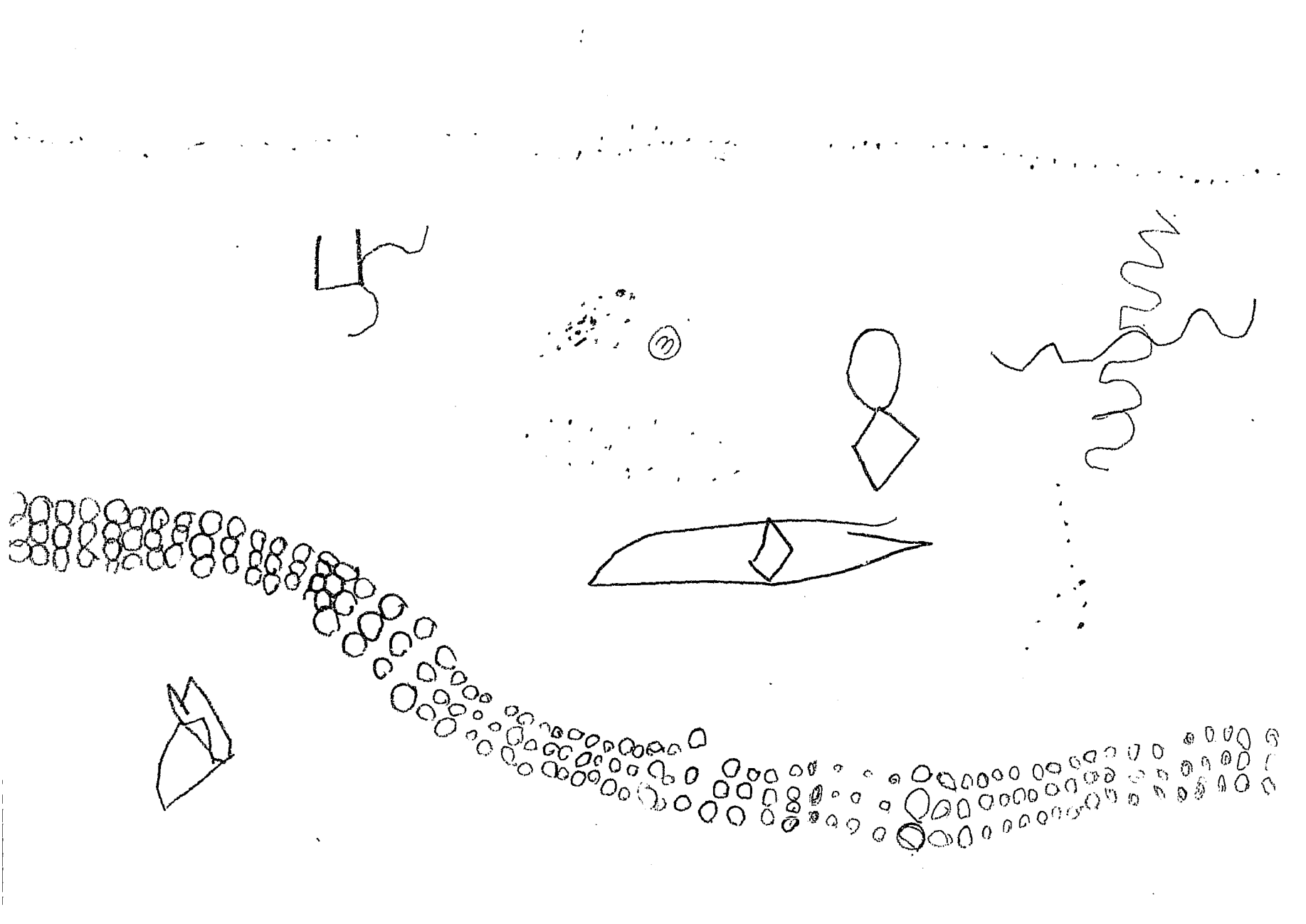
Please tick only **one** of the following:

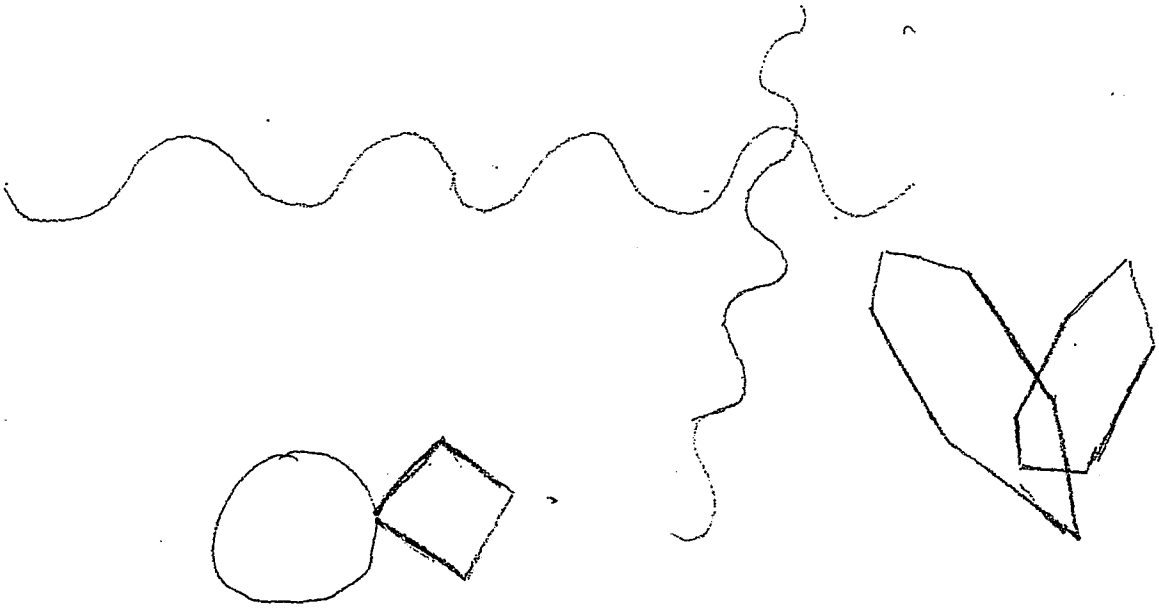
1. Primarily exhibiting emotional difficulties ___
2. Primarily exhibiting visual-perceptual difficulties ___
3. Exhibiting both difficulties ___
4. Exhibiting neither difficulty ✓

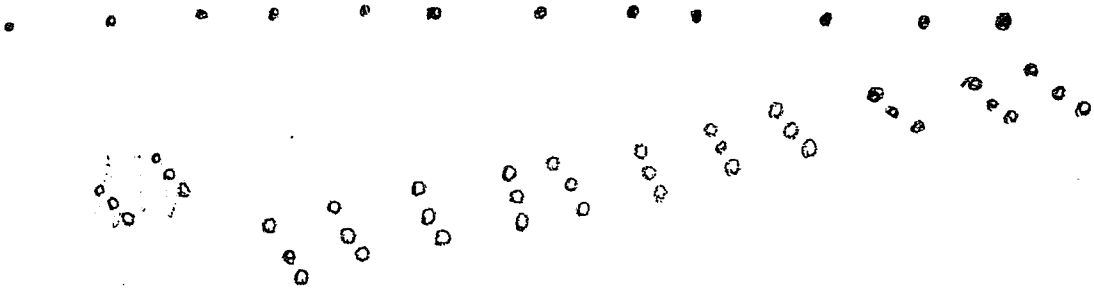

23
23

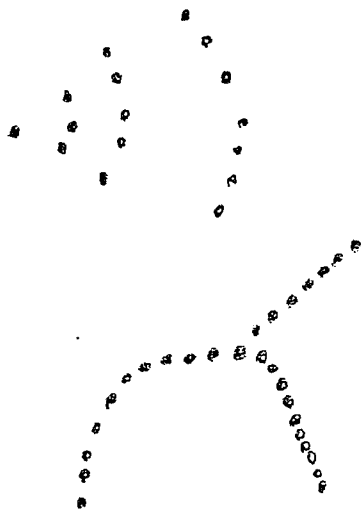


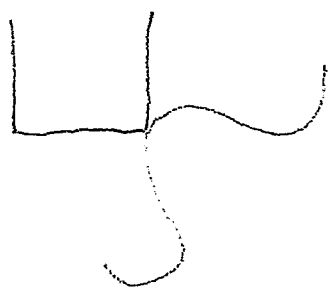


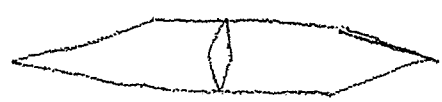


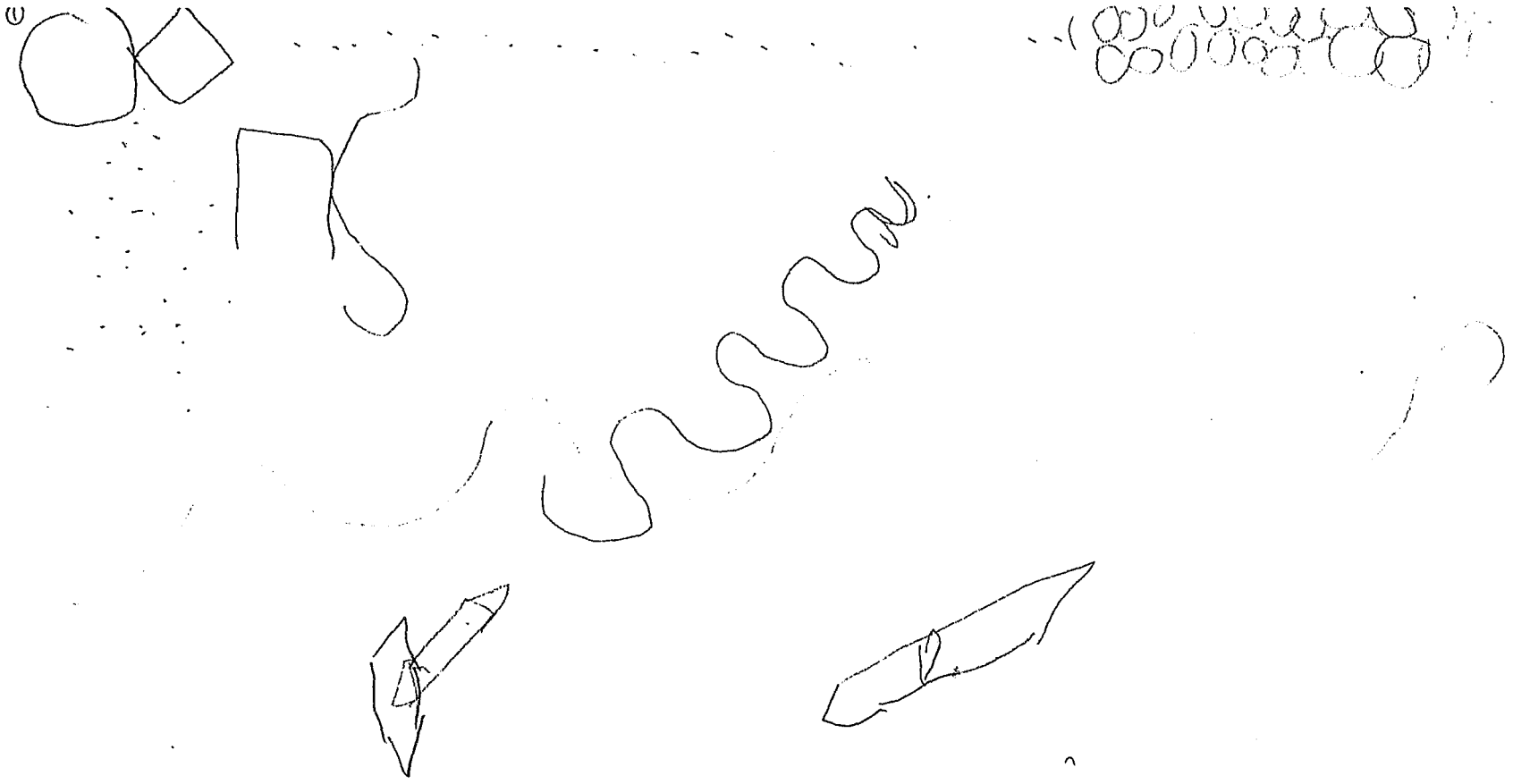












Author Suttner V
Name of thesis Inter And Intra-Rater Reliability Of The Scoring And Analysis Of Bender Gestalt And Human Figure Drawing Protocols In A Clinical Setting Suttner V 2000

PUBLISHER:
University of the Witwatersrand, Johannesburg
©2013

LEGAL NOTICES:

Copyright Notice: All materials on the University of the Witwatersrand, Johannesburg Library website are protected by South African copyright law and may not be distributed, transmitted, displayed, or otherwise published in any format, without the prior written permission of the copyright owner.

Disclaimer and Terms of Use: Provided that you maintain all copyright and other notices contained therein, you may download material (one machine readable copy and one print copy per page) for your personal and/or educational non-commercial use only.

The University of the Witwatersrand, Johannesburg, is not responsible for any errors or omissions and excludes any and all liability for any errors in or omissions from the information on the Library website.