Psychological Testing

American Academy of Matrimonial Lawyers
March 9, 2005
St. Thomas

By
Mary Johanna M^c^curley
M^c^Curley, Orsinger, M^c^Curley, Nelson & Downing, L.L.P.
5950 Sherry Lane, Suite 800
Dallas, Texas 75225
Phone: 214/273-2400

Kathryn J. Murphy
Koons, Fuller, Vanden Eykel & Robertson
A PROFESSIONAL CORPORATION
5700 W. Plano Pkwy., Suite 2200
Plano, Texas 75093
(972) 769-2727
Internet: Kathryn@koonsfuller.com

I. INTRODUCTION

The family law practitioner is confronted with the results of psychological testing with increasing frequency. The tests are generally interpreted as establishing the mental health and parenting skills of a parent.

Psychological testing became popular during World War II, and clinical psychologists have been toting test kits ever since. As this article illustrates, psychological tests have a potentially valuable role to play in the assessment of the mental health of parents and of children. However, more problematic is the application of psychological testing to the resolution of custody issues. Because all psychological tests are eventually filtered through the psychologist's interpretation, the tests can be no better or worse than the interpretation itself.

This article supports the continued development, administration and use of psychological testing in custody disputes, but it also sends a warning to attorneys to not always have faith that these tests are valid, fair, or even constitutionally appropriate. As David Viscount, in his book, The Making of a Psychiatrist, observes, "Mental health professionals love their interpretations, they are gifts that they give themselves."

II. GENERAL CONSIDERATIONS OF PSYCHOLOGICAL TESTING

Psychological testing is a standardized method of checking a portion of an individual's behavior and comparing it to that of a group with known characteristics. The tests are categorized depending on what factors the particular test is designed to measure. Intelligence tests, personality tests, achievement tests, and neuropsychological tests have all been developed in an effort to create objective and standardized instruments. [D. Shuman, Psychiatric and Psychological Evidence 47 (1986).] Since cases involving the parent-child relationship involve psychological issues and, therefore, often require psychological testimony regarding intelligence and personality, this article will examine the major tests in these areas.

In appraising a particular psychological test, there are two considerations which must always be kept in mind: **reliability** and **validity** of the particular test.

A. Reliability

Reliability of a given test requires consistency of results. For example, in order for a test to be "reliable" in scientific terms, the test results should be essentially the same whether the patient is tested six days, six weeks or six months later.

The legal analogy of reliability is witness credibility. If Mr. Jones says in a deposition that Mrs. Smith is a good mother, but at trial he says she is not always a good mother, then he is an "unreliable" witness. What was true at the deposition should also be true at trial. If someone is tested on August 1, 2004 and again on January 1, 2005, then the test results should be essentially the same if the test is reliable.

Reliability of a test is usually determined in terms of a coefficient correlation between test scores on a first testing and then on a second. A reliability coefficient value ranges from 0.0 (no reliability) to +1.0 (perfect reliability). There is not complete agreement among psychologists as to what reliability standard should be set in order to be confident of a certain test. "Correlation coefficient" means, in statistical terms, the ratio between two different quantitative measures, in this situation representing the degree of association of the variable phenomena of a first test taking versus a second. In laymen's terms, this means that the closer the results are between the first testing and the second testing, the closer to the coefficient of a 1.0 the results will be. The more conservative group believes a correlation coefficient of 0.90 is the lowest acceptable correlation coefficient for a test to be considered "reliable." Another school of thought argues that tests with a reliability coefficient of 0.80 are sufficient. Clearly, a test with less than a 0.80 reliability coefficient does not meet the standards of reliability accepted by either school of thought. [J. Ziskin, Coping with Psychiatric and Psychological Testimony, Vol. I, 211 (1981).] However, the attorney should be aware that for every psychologist and psychiatrist existing, there is a different opinion as to what the lower limits of reliability ought to be.

B. Validity

Validation of a test, for scientific purposes, requires that the test measure what it actually purports to measure. For example, does a particular intelligence test truly measure intelligence? The American Psychological Association, along with two other associations, have produced a report entitled, "Educational and Psychological Tests and Manuals, 1974." The report is designed to help the clinical psychologist determine "validity" of a certain test.

However, validity is more difficult to establish than reliability. For example, assume five witnesses all testify that Mr. Jones is a good businessman (high reliability). Is there relevance to that undisputed testimony, when the issue is whether he is cruel to his wife? In other words, no valid relationship has been established between being a good businessman and mental cruelty.

There are different approaches to validity, which are:

1. "PREDICTIVE" VALIDITY - The most commonly used and useful validation process is predictive validity. This means that with certain information about A, one can state with probability that B will occur.(Id. at 78.)

2. "CONCURRENT" VALIDITY - Concurrent validity is essentially the same as predictive validity, except instead of being predictive of behavior, the test scores are compared to present information. Comparing what score one makes on an intelligence test with what grades the test taker is making in school is an example of concurrent validity. (Id. at 80.)

3. "CONTENT" OR "FACE" VALIDITY - This kind of validity is based on what we call in law, res ipsa loquitur, "the thing speaks for itself." If the test involved problems in multiplication, it would then seem to follow that the test measures one's ability to multiply.

4. "CONSTRUCT" VALIDITY - This is an all-encompassing concept which assimilates one or more of the other types of validity. "It refers to the extent to which the test may be said to measure a theoretical concept, for example, intelligence or mechanical comprehension or anxiety. It entails a broader, more enduring and more abstract kind of behavioral description." (Id. at 81.)

III. INTELLIGENCE TESTS

A. Purpose of Intelligence Tests.

Intelligence tests are a standardized method of evaluating mental ability, which aids the clinical psychologist in the diagnosis of individuals. Intelligence tests are used to achieve four main purposes:

1. An Appraisal of the Intellectual or Mental Capacity of an Individual.

2. Indications of Possible Personality Disturbance.

An intelligence test may indicate a personality disturbance when the test taker's performance is erratic. Such erratic performance is evident when the test taker's responses to questions are not consistent. For example, the test taker answers the easier questions incorrectly, but is able to answer the more difficult questions correctly. Additionally, if the test taker gives strange answers to questions, this may be indicative of a serious mental disturbance.

Intelligence tests can never be used alone to diagnose a patient, although they can contribute to the understanding of the patient's behavior.

3. Indications of Special Abilities or Limitations

A low verbal score may be evidence of lack of education or it may mean the loss or impairment of the power to use words as symbols as the result of some brain injury. Likewise, the test may reveal an unusually high ability in the use of vocabulary or arithmetical computation. Depending on the presenting problem to the psychologist, such information can add to the understanding of the patient's skills and potentialities.

4. Observations of the Subject's Behavior

Often, personality trends may be picked up by close observation during the testing. For example, does the testing make the test taker ill, does it make him or her tense, or make him or her at ease? [S. Garfield, Clinical Psychology 113 (1983).]

One of the major problems with intelligence tests is that they tend to include an abundance of verbal testing, and they are also biased toward classes of individuals with particular educational and cultural opportunities. Additionally, as stated earlier, if the person administering the test does not follow the standard instructions and methods of administration designed for that particular test, then the results will be faulty. If the test is not given and scored as directed by the test manual, then it cannot be interpreted in terms of the existing norms. Even more alarming is that sometimes the standard instructions are incomplete or inadequate and the clinician must rely on his or her own judgment, thus making the test only as good as the clinician and his or her judgment. (S. Garfield, supra, at 113).

The two most widely used intelligence tests are the revised Stanford-Binet and the various Wechsler Scales. (Id. at 121.)

B. Stanford-Binet Intelligence Scale

1. History and Construction of the Stanford-Binet Test

The Stanford-Binet Intelligence Scale is designed to measure cognitive abilities in anyone over the age of two. It is used to analyze patterns of thinking, as well as overall cognitive development. The test consists of subparts that measure different areas of cognitive development. These subparts include Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory. The tests will produce a Composite Standard Age Score which is referred to as "SAS". (Buros Desk Reference, Psychological Assessment in the Schools (1994). The test consists of 15 subsections, which are not administered to everyone. The examinee's age and performance will determine the number of subtests that are given.

The Stanford-Binet was originally published in 1916, and revised in 1937, 1960 and 1972. It is currently in its 4th Edition. It was devised as a measure of children's intelligence and

later revised to include adults. (D. Shuman, supra note 1, at 51.) The scale is based on mental age. Mental age is determined by groups of test items which are arranged in terms of age levels. If the patient is below the age of six, the items are grouped into half year levels (for example: 4, 4-1/2, 5,5-1/2, etc.). If the patient is six to fourteen years of age, the items are grouped into year levels of 6, 7, 8, and so on. After the age of fourteen, there are four levels ranging from "Average Adult" to "Superior Adult, III." (S. Garfield, supra, at 113).

At the adult levels, verbal testing is mostly used, as well as problem solving and abstract thinking; whereas, testing at the children's level uses blocks and toys, as well as verbal testing. (Id. at 122.) The test is administered by beginning at the level of the person's chronological age, or just below it, and then proceeding up a level until the person fails all the items on that level, or going down from the initial level if the person answers all of the items on that level incorrectly. (Id. at 122.) There is no administration time reported for the entire testing, although the Pattern Analysis subtest does have strict time limits. The test is convenient to use. The flip-over test booklets have the directions on the examiners side, and the display for the examinee on the other.

The I.Q. (intelligence quotient) is an index which indicates the rate of mental development. I.Q. is determined by dividing the determined mental age by the individual's chronological age and multiplying by 100. For example, a child who is ten and obtains mental age of ten, would have an I.Q. of 100. If the same child secured a mental level of twelve, he or she would receive an I.Q. of 120. (Id. at 123.) However, the Stanford-Binet has adopted the deviation approach of a mean of 100 and a standard deviation of 16. [C. Golden, Clinical Interpretation of Objective Psychological Tests 2 (1979).]

Approximately fifty percent of the people who take the Stanford-Binet obtain a score of between 90 and 110, twenty-five percent achieve above 110, and twenty-five percent below 90. Those who score above 140 represent one percent of the population.

The following shows examples of what is on the test range at the nine year level:

1. Paper Cutting – The child is asked to cut out six inch squares of paper which have been folded and then is asked to make a drawing of how the paper would look unfolded.

2. Verbal Absurdities – A series of "foolish" statements are read and the child must tell the examiner what is foolish about the statement.

3. Memory for Designs – A card with two designs is shown to the child and then the child is asked to draw the designs by memory.
4. Rhymes – The child is to tell the name of a color, a number, an animal, or a flower which rhymes with a specific word.

5. Making Change – The child is simply required to make change. For example, I give you a dollar for something which costs fifty cents. How much change do I get back?

6. Repeating Four Digits Reversed – The child must repeat, backwards, a series of four digits given to him or her. (S. Garfield, supra note 6, at 124.)

On the Superior Adult III level, the test taker must be able to define thirty words correctly, such as ocher, incrustation and perfunctory. (Id. at 124.)

The 1960 revision was based on testing about 4,500 subjects from ages 2-1/2 to 18 years. Only those testing items which proved to be the best were used in the revision. The reliability standard for the items are generally in the 0.90's. In the 1972 revision, about 2,100 cases were tested, and this sample was even more representative of the general population than the original. (Id. at 125.)

2. Reliability/validity

There are at least two limitations to the Stanford-Binet scale. First of all, the test is clearly dominated by testing verbal abilities (except at the very earliest ages) and is, thus, an inadequate test to test intelligence for a person with limited verbal capabilities, whether from a lack of education or a different cultural upbringing. With the growing number of Asians and Mexicans in our culture, for example, this problem with the test should be kept in mind. Secondly, persons over the age of 18 were not represented in the standardized sample, and the test is, therefore, not particularly well suited for adults. Additionally, some of the actual test items relate mainly to children or to school. (Id. at 125.) The test may very well be only appropriate for children from ages 5 to 15 or for suspected retarded adults. (Id. at 126.)

The Stanford-Binet's predictive validity has ranged from correlation coefficients of 0.40 to 0.75 when compared to school grades, teacher ratings and achievement test scores. The usefulness of the test for any other purposes has never been adequately demonstrated. (Ziskin, supra, at 215.)

Additionally, evidence indicates that the IQ scores can be affected by such things as different examiners, home environment, and the present health of the individual. (Id. at 215.) Obviously, a custody fight affects the home environment and sometimes even the health of an individual. The clinician who has given a Stanford-Binet in a custody case probably cannot assure the court of its validity, so it is the duty of the attorney on cross-examination to explore this.

C. Wechsler Adult Intelligence Scale - Revised or Third Edition (WAIS-R)

1. History and Construction of the Wais-r

The Wechsler scale was developed in 1939, but has been updated several times since then, the latest in 1981. (Id. at 127.) This test was devised specifically to meet the needs of a standardized test for adults. The newest version is the Wechsler Adult Intelligence Scale - Revised (WAIS-R) and is now considered the most widely used adult intelligence test in the

United States. [Lubin, Larson & Matarazzo, "Patterns of Psychological Test Usage in the United States: 1935 - 1982," 39 Am Psychologist 452-453 (1984).]

Not only are the Wechsler Scales different from the Stanford-Binet in that they were developed for adults, the Wechsler Scales are point scales rather than mental age scales. Points are given for correct answers, and then those points are converted into standard scores and then to an I.Q. Another difference between the Wechsler Scales and the Stanford-Binet is that where a certain type of test item is found throughout the Stanford-Binet at different levels, the Wechsler Scales group like items together. For example, on the Stanford-Binet, a 2-1/2 year old might be requested to repeat two digits, and on each new level, the difficulty is increased. On the Wechsler Scales, all of the digit items are grouped together as one sub-test and administered from the easiest to the most difficult.

The Wechsler scales also have age norms for adults. It appears from normative samples that adults reach the peak of their mental development between the ages of 25 and 34. After 34, there is a slow decline. For example, according to the WAIS-R manual, a score of 115 at ages 25 - 34 is equal to an I.Q. of 100. At ages 34 - 44, the same score is equal to an I.Q. of 105; and at ages 55 - 64, it is equivalent to an I.Q. of 113; and from 70 - 74, equal to an I.Q. of 122. (S. Garfield, supra note 6, at 127-29.)

The test has two major categories of tasks: verbal and performance. Each category has sub-tests which measure different abilities, enabling the clinician to analyze the pattern of scores across the sub-tests, as well as verbal performance and full scale I.Q.s. This testing method assists the clinician in the diagnosis of psychiatric disorders, chronic alcoholism, and brain damage. [C. Golden, supra.]

a. Sub-tests

The WAIS-R consists of eleven sub-tests. Six sub-tests are in the verbal category and five are in the performance category. Items are arranged in each sub-test from the simplest to the most difficult. [D. Wechsler, WAIS-R Manual: Wechsler Adult Intelligence Scale - Revised (1981).]

Each sub-test appears below in the order in which it appears in the test. Note that the verbal sub-tests are not all grouped together, nor are the performance sub-tests, but are rather interspersed throughout the test:

Information - Verbal
Picture Completion - Performance
Digit Span - Verbal
Picture Arrangement - Performance
Vocabulary - Verbal
Block Design - Performance
Arithmetic - Verbal
Object Assembly - Performance
Comprehension - Verbal
Digit Symbol - Performance

Similarities - Verbal

b. VERBAL SUB-TEST

(a) Information - The information sub-test consists of items to test the test taker's general knowledge of information of specific facts. Low scores on this sub-test are associated with low intelligence. (C. Golden, supra note 90, at 18.)

(b) Digit Span - The digit span sub-test is a test of immediate memory. Scores are given for memory of digits, forward and backward, for a combined score. (Id. at 18.)

(c) Vocabulary - The Vocabulary sub-test is considered the best estimate of a person's intelligence. High scores are made by those with obviously high verbal intelligence and advanced education. (Id. at 19.)

(d) Arithmetic - The arithmetic sub-test, which is interestingly in the verbal category, measures an individual's ability to work number concepts logically and in the context of daily problems. (Id. at 16.)

(e) Comprehension - The comprehension sub-test is a measure of the test taker's ability to understand social customs and to show the appropriate response in given situations, as well as the reasons for particular responses. The sub-test obviously assesses the individual's socialization and assimilation in society, meaning how well he relates and interacts with the rest of society. The test requires the individual to select, from alternative answers, the most logical answer. Additionally, the test evaluates the test taker's long-term memory and experiences. (Id. at 15.)

One advantage of this particular sub-test is that it elicits remarks from the individuals which may reveal a pathological condition, if the remarks are bizarre or unusually odd. For example, if an individual is asked why we register marriages, and he answers, "So that the government can keep track of children who are to be sent to top secret camps," then the clinician would have a clue that this individual may have a problem. This may even offer diagnostic clues. (Id. at 15-6.) Good scores on this sub-test is a good indication of whether someone is in touch with reality. (Id. at 16.)

(f) Similarities - The last sub-test of the verbal category is the similarities sub-test. The individual is asked to tell how two objects are alike. For example, out of "tools" or "clothing" or "horses," which two are the most similar? The more abstract the person's association, the higher the score he or she will receive on the sub-test. More concrete thinking is usually related to the low test scores. (Id. at 17.) Generally, it is thought that schizophrenics tend to deny the presence of similarities. (Id. at 18.)

c. PERFORMANCE SUB-TEST

Performance sub-tests are designed to be more dependent on visual, spatial, or sequential abilities rather than on verbal skills.

(a) Picture Completion - The picture completion sub-test consists of a series of drawings or sketches where an essential element is missing. The picture completion sub-test is often the best estimate of what is called "pre-morbid intelligence," a term used to refer to one's intelligence prior to the occurrence of a head injury or other mental illness or condition. It appears that paranoid patients often believe that nothing is missing from the picture. Weiner (1966) believes that schizophrenia can be determined by what errors are made. For example, a schizophrenic might find a tongue or sex organs missing on a picture of a horse with a missing tail. Or, on a violin that is missing strings, the schizophrenic might say the violinist is missing. (Id. at 20.)

(b) Picture Arrangement - The picture arrangement sub-test is designed to evaluate a person's perception of certain pictures, his or her organization of several pictures, awareness of appropriate social sequences, planning skills, ability to form and test hypotheses, and flexibility and ability to sequence items in a logical order. Wechsler called this test (1958) a measure of "social intelligence." (Id. at 21.)

(c) Block Design - The block design sub-test is the purest measure of nonverbal reasoning in the Wechsler Scale. It is, therefore, the best measure of nonverbal intelligence and general spatial skills. For example, the skills necessary in basic engineering may be tested. The test taker views the block design that the examiner has made and then attempts to reconstruct it. Obviously, the test requires visual analysis skills and visual motor coordination. The test allows the psychologist to observe how the subject goes about solving problems, whether it is impulsively or in an organized manner. This sub-test is a good measure of differentiating between schizophrenics and brain-damaged patients. (C. Golden, supra note 14, at 21.)

(d) Object Assembly - The object assembly sub-test is not unlike the block design sub-test in that it requires visual motor skills. It is unlike the block design, in that rather than forming abstract designs, the individual must put together puzzles of familiar objects. For a high score, speed and accuracy are essential. Rapaport, Gill & Schafter believe that depressed individuals tend to make low scores on this sub-test. (Id. at 22.)

(e) Digit Symbol - The digit symbol sub-test is sensitive to motor problems in the dominant hand and measures basic learning skills. The test taker must also associate a symbol with a number. A good visual memory is required as well as quickness. This test is thought to be a very good measure of high levels of anxiety. (Id. at 19-20.) Lower scores are often indications of conditions ranging from anxiety to brain damage. This sub-test is well suited to being given to the culturally deprived or uneducated because it relies very little on verbal skills.

2. Reliability/Validity of the Wais-R

The overall reliability of a total score achieved on the WAIS is 0.97; however, the WAIS has several sub-tests which have reliability scores as low as 0.60. An expert witness who has given the WAIS may conclude that it may not reach acceptable reliability limits.

a. Verbal Sub-tests

(a) Information. Although a low information sub-test score may mean low intelligence, it can also be the result of the subject being from a different cultural background than the norm group or having a poor educational background or mental disorder. (C. Golden, supra, at 14.)

(b) Digit Span. The digit span sub-test, although basically a test of memory, can be affected by a lack of concentration or attention. Thus, if the test taker is distracted because of the problems involved in his or her divorce or custody suit, then the test results may not be accurate. (Shuman, supra, at 49.)

(c) Vocabulary. The vocabulary sub-test, although considered the best estimate of intelligence on the WAIS, is also the most adversely affected by differences in cultural and socio-economic background. (Id. at 49.)

(d) Arithmetic. The biggest flaw with the arithmetic sub-test is that anyone who is distracted or highly anxious may have low scores. Once again, so many of the persons who are in custody litigation are naturally distracted and anxious. Additionally, it is reported that many individuals react poorly and even refused to answer the problems presented them because they feel the problems are stupid. Thus, an otherwise bright individual may end up with a low score. Furthermore, this sub-test may not be a particularly good indication of arithmetic skills due to the non-use of pencil and paper on this sub-test; and further due to the emphasis placed on memory, concentration, attention and verbal skills. (C. Golden, supra, at 16-7.)

(e) Comprehension. While a good score on the comprehensive sub-test is a good indication that the individual has a good grasp of social realties, studies have shown that the opposite is not necessarily true if the individual has a low score on this sub-test. (Id. at 16.)

(f) Similarities. Although it is generally thought that in the similarities sub-test, schizophrenics deny the presence of similarities, it is rare that the answers on this sub-test are indicative of a specific pathology. (Id. at 18.)

b. Performance Sub-tests

It is generally assumed that the performance sub-tests and the verbal sub-tests measure different things. However, this assumption appears to be false. Almost every performance test involves verbal abilities. For example, the Digit Symbol sub-test incorporates numbers and symbols, not unlike letters, and it, therefore, offers a clear advantage to the literate or educated individual. The picture completion sub-test, on the other hand, can only be understood through verbal instruction. (Id. at 28-9.)

(a) Picture Completion. There is a strong cultural component to the task associated with the picture completion sub-tests because familiarity with the object pictured is necessary for high scores on this sub-test. For example, someone who has grown up in Harlem and has never left might not be able to complete the stirrups on a saddle. (Id. at 20.)

(b) Picture Arrangement. Once again, the biggest drawback to this sub-test is the component of being from a particular culture. (Id. at 21.)

(c) Block Design. The biggest disadvantage of the block design sub-test is the difficulty that low intelligence individuals may have taking it because there are not easily answerable items on this test. (Id. at 21.)

(d) Object Assembly. It has been suggested that depressed individuals do poorly on the object assembly sub-test because of the heavy emphasis on time bonuses used in scoring this sub-test. (Id. at 22.) The attorney might want to ask the witness who has administered this test if situational depression in your client might affect his or her score on this test.

(e) Digit Symbol. Because of the role that visual acuity, motor coordination and speed play in this sub-test, it may be a more difficult sub-test for older adults. Studies have shown that older adults do not write or handle objects as quickly as younger adults. Thus, there is some question as to whether speed should be given any weight in the evaluation of the individual's intelligence. [Matarazzo, Wechsler's Measurement and Appraisal of Adult Intelligence, 215 (1972).]

3. Advantages of the Wais-R

There are a number of advantages in using the WAIS-R as a measure of intelligence:

a. The WAIS-R is a standard against which all other I.Q. tests may currently be measured.
b. The WAIS-R does not just give the clinician a final score, but also sub-test scores, which can give important information which an overall score cannot give.
c. It is presently the most comprehensive normed adult intelligence test available.
d. The WAIS-R has been the most heavily researched test available to the psychologist. C. Golden, supra, at 33.)

4. Disadvantages of the Wais-R

a. The WAIS-R is not suitable for large group testings increases the cost of its administration.
b. Since the WAIS is biased toward the average American culture group, there are problems administering it to Blacks, Mexicans and other minority groups. (There is an intelligence test called BITCH,

which is the Black Intelligence Test Corrected for Honkies, which is a test specifically designed for black Americans and which is suggested by some researchers to be used in place of the WAIS when the individual to be tested is black).

c. The WAIS has a tendency to produce higher scores on retesting (as do most intelligence tests).

d. The WAIS has the tendency to overestimate low I.Q.s. A person with a score of 0 might actually be an even lower I.Q. (Id. at 33-4.)

e. The most serious problem of the WAIS is the tendency of clinicians to use the test for diagnostic uses, not merely as a tool to measure intelligence. The scientific validity requires that the test measure what it actually purports to measure. For example, Wechsler (1944) and Rabin (1941) reported that their studies showed that schizophrenics obtain higher scores on the Verbal Scale than on the performance Scale. These findings by Wechsler and Rabin have not, however, been confirmed by other studies. Similar contradictory findings of other diagnostic uses of the WAIS have also been reported. Many scientific investigators have called attention to the conflicting results and patterns reported for various clinical populations (Garfield 1948, 1949; Carter and Bowles, 1948; Guertin, Rabin, Frank and Ladd, 1962; Harper, 1950; Hunt and Cofer, 1944; Rabin 1945; Rabin and Guertin, 1941). (Garfield, supra, at 130 -31.)

Despite these findings, psychologists have continued to use the WAIS for diagnostic uses. Since the revision of the WAIS, more recent studies indicate some improvement in the quality of research on using the WAIS for diagnostic purposes. However, the research still does not reveal truly reliable or valid patterns of opinions on using the WAIS to diagnose. (Guertin, et al., 1966, 1971). (Id. at 131.)

G. H. Frank wrote the following, in 1970, concerning the use of Wechsler Scales as diagnostic tools:

> "The overview of the more than twenty-five years of research, therefore, presents us with studies that are inconsistent, contradictory, and hence, inconclusive. What does seem clear, however, is that the specific predictions regarding the performance of the suggests in the major diagnostic categories, viz., schizophrenia, neurosis, or the brain-disordered, as postulated by Wechsler and Rapport, have not received support. Indeed, the research does reveal that there is no characteristic pattern of performance on the sub-tests for, for example, the schizophrenic or the neurotic (page 177). (Id. at 132.)

Any cross-examination of an expert witness who has used Wechsler Scales in his or her evaluation of an individual should include inquiries into how the WAIS was used in his or her evaluation.

V. PERSONALITY TESTS

While intelligence tests are designed to evaluate the intellectual ability of an individual, the purpose of personality tests is to evaluate different aspects of a person's emotional and social functioning. There are two types of personality tests - objective and projective. The following covers the major and most used objective and projective personality tests.

A. Objective Personality Tests

Objective personality tests present to the subject test taker questions or statements which he or she is required to answer by choosing among a group of alternative answers. The choice of answers may be "true-false," "sometimes-always-never," or "agree-disagree." The most commonly used objective personality test is the Minnesota Multiphasic Personality Inventory (MMPI). (Shuman, supra note 1, at 51.)

1. Minnesota Multiphasic Personality Inventory (Mmpi)

a. History and Background of MMPI

The MMPI, which was developed in the 1930's and first published in 1943, has become the "premier diagnostic and screening devise in clinical psychology." (Id. at 54.) Stark Hathaway, Ph. D., and John McKinley, M.D., Ph.D., the MMPI's creators, sought to develop a tool to condense psychiatric interviews which were both lengthy and very expensive to administer to the mental health patient. The MMPI's basic purpose is that of a diagnostic tool to examine psychological pathology. Much as an x-ray machine looks inside the body, the MMPI seeks insight into an individual's psyche. (M. McCurley and K. Fuller, "The MMPI - What is It?," Advanced Family Law Course, 1986, State Bar of Texas.) Although psychiatric diagnosis is far more difficult than physical diagnosis, the MMPI has been used in so many settings for over forty years that an enormous body of literature has resulted regarding diagnostic and treatment implications of various MMPI profiles. (Id. at 54.)

b. General Terms and Definitions

The following is a list of key terms used in the application of the MMPI.

(1) "MMPI" - As indicated above, the initials stand for Minnesota Personality Inventory. The test is known as a standardized, objective test; however, there are several differing "flavors" of the MMPI. The standard MMPI most commonly seen in custody litigation consists of 566 questions. Some of the variations of the standard 566 question MMPI include the MMPI-168 (which contains only 168 questions), the Mini-Mult (71 questions), the Midi-Mult (86 questions), Fashingbauer's Abbreviated MMPI (commonly known as the FAM and containing 166 questions), and the Maxi-Mult (94 questions), and the MMPI-2 (which will be discussed later in this paper). It should be noted that currently the various short forms described above are sanctioned by the University of Minnesota, which publishes the

MMPI for use as a research tool only and not for clinical use. (McCurley and Fuller, supra note 40, at T-2.)

As seen from the outset, defining the MMPI can be difficult because the term may be used generically by psychologists to describe any of the variations of the test. The MMPI most commonly administered to individuals for individual evaluation, as opposed to group evaluation, is the standard 566 question MMPI. (Id. at T-2.)

(2) "ITEM" - This is the term used to refer to the questions appearing in the MMPI. The items are not really questions, however. Each item is an affirmative statement written in the first person singular, such as item #231 in the original MMPI, which reads, "I like to talk about sex." To each of the items, the respondent is instructed to answer "true" if the statement is "mostly true" or "false" if the statement is "mostly false," as the statement applies to the respondent. If the respondent cannot answer "true" or "false" to the items, the respondent is allowed to answer "cannot say." The items are presented to the respondent usually in the form of a test booklet and the respondent fills in his or her answers on an answer form. (Id. at T-2.)

(3) "SCALES" - The standard MMPI is broken down into 14 categories known as "scales." Four of the scales are devoted to test validity; i.e., are the responses of the test taker answered in such a manner that the test is subject to a valid interpretation of that person's personality? The remaining ten scales are "clinical scales." The clinical scales are designed to measure the psychological inventory of the respondent. The standard MMPI, as previously mentioned, contains ten clinical scales; however, many versions of the MMPI contain more than the original ten scales. It should be understood at the outset that the items appearing on the MMPI test form are not arranged by scale. After the 566 items which make up the MMPI were selected, the scales were developed by determining which group or groupings of items and their answers indicated certain personality traits sought to be measured. A "Scale" on the original MMPI is simply that group of items and their responses which purport to measure a phenomena. For example, part of Scale O (Si) (70 items) is made up of items 32, 67, 82, 111, 117, 124 and so on. (Id. at T-3.)

(4) "T" SCORES AND "RAW" SCORES - A "raw" score on the MMPI indicates how many items in the particular scale the respondent answered in the "critical direction." The "critical direction" means the item was answered in such a manner to elevate the score; i.e., if a respondent answered the item, "I do not like everyone I know,", "false," this would elevate the score on the "L" scale. The "T" score is the eventual score shown on the MMPI score sheet. Normal "T" scores generally range from 30 to 70 on each scale. The T-score is arrived at by taking the raw scores through various statistical calculations which are not easily understandable without a sophisticated knowledge of statistics. (Id. at T-3.)

(5) MALE/FEMALE SCALES - The design of the MMPI recognizes that males and females may, and in some circumstances should, respond differently to certain items appearing on the MMPI. Therefore, males and females are scored differently, and the score sheet will usually so indicate by denoting MMPI-Male or MMPI-Female. (Id. at T-3.) A special warning to lawyers on this scale. The Male/Female Scales are only a measure of those

characteristics that have been observed as typical, traditional, conservative male and female values. So, highly educated males (lawyers, psychologists, doctors, etc.), will score higher on the female scale because of their typical appreciation of art and other aesthetic values. Likewise, lower educated females will score higher on the masculine scale because of their typical cultural association with more masculine environments (drag races, tractor pulls, etc.). Thus, the lawyer should not be too quick to use this scale to indict the highly educated males as homosexuals or the lower-educated females as lesbians (which many divorce lawyers are apt to do in cross-examination). This, like all other scales, must be considered in the social context of the subject of study.

2. Construction of the MMPI

a. STEP 1: SELECTING THE ITEMS: Hathaway and McKinley gathered over 1,000 questions or items that had long been in use in the psychological texts and psychiatric examination forms of their time (circa 1940). After reviewing various research works and after exercising their subjective and professional opinions, Hathaway and McKinley, narrowed the item pool to 550. (Remember, 16 items are repeated.) (Id. at T-4.)

According to the creators of the MMPI, no item was eliminated from the final pool of items only because its manifest content seemed to have no relation to the psychological syndrome in question. However, they claim that no item was arbitrarily included in the final pool of items if the validating evidence for the items was not strong. Hathaway, S.R., Basic Readings on the MMPI in Psychology and Medicine, University of Minnesota Press, Minneapolis, 1956, p. 106. Put another way, the creators of the MMPI did not throw out any items merely because the wording of the item did not seem, on it face, to have any relevance to the particular personality trait being measured. For example, one question, purportedly a potential indicator of depression, reads, "I like to flirt." Although to the layman this statement may not, on its face, seem to have anything to do with depression, according to the rationale of Hathaway and McKinley, there was prior validating evidence that a false answer to that statement is an indicator that a person may be depressed. (Id. at T-5.)

b. STEP 2: GROUPING THE ITEMS INTO GENERAL CATEGORIES: As the items were being selected, they were divided into 26 different categories. These categories included topics ranging from "General Health" to "Habits" to "Obsessive Compulsive" behavior.

c. STEP 3: SELECTION OF THE NORMATIVE SAMPLE GROUP: The scoring and interpretation of the MMPI is essentially based on comparing the score of the individual taking the test with that of a sample group of psychologically normal and abnormal people; i.e., did the person answer the questions like the normal folks or like the "crazy" ones.

The character and attributes of the normative group are very important because the character and attributes of the group comprise the cornerstone of the original MMPI. One researcher described the importance of the MMPI normative group as follows:

"... the performance of these men and women on each of the component scales in the MMPI is used as the basis for the norms in the test profile. Each subject taking the MMPI, therefore, is being compared to the way a typical man or woman endorsed those items. In 1940, such a Minnesota normal adult was about thirty-five years old, married, lived in a small town or rural area, had eight years of general schooling, and worked at a skilled or semi-skilled trade (or was married to a man with such an occupation level)." Dahlstrom, Welsh, and Dahlstrom, An MMPI Handbook, Vol. 1: Clinical Interpretation. University of Minnesota Press, Minneapolis, 1972, pages 7 - 8. See Colligan, 1983, pages 67 - 68, for a detailed discussion regarding the discrepancies between the original 1940 group and its refined subset here identified as the 1957 group.

d. STEP 4: CREATING THE SCALES: As stated earlier, the original MMPI has 14 scales (four validity scales and 10 clinical scales). A basic understanding of how theses scales were developed is essential to understanding the MMPI. As discussed above in Step 3, normal sample groups were selected. Additionally, sample groups of individuals diagnosed as having various psychological ailments, such as severe depression, hypochondria, and schizophrenia, were selected. Essentially, the clinical scales were established by giving the MMPI to a group of normal respondents and a group of, for example, schizophrenics. The scale for schizophrenics would then be developed by comparing how the two groups answered various items. For example, the scale of the original MMPI denoted Scale 8 (Sc), (schizophrenia was essentially developed by noting which items the schizophrenics answered in a true or false manner and comparing those responses to the responses of the normal sample and grouping the items commonly endorsed by the schizophrenics into Scale 8 (sc). (Id. at T-6.)

3. Reliability and Validity of the MMPI

If the MMPI were a perfect instrument for inventorying a person's psychological traits, one would expect that unless the individual goes through a major change in personality, multiple MMPIs given to the individual would produce similar results. Research has been done by Roger Green attempting to examine "test and retest" reliability for the MMPI.

Additionally, according to reviewers in Buros Fifth Mental Measurements Yearbook, [O. K. Buros, Fifth Mental Measurements Yearbook (1959)] the validity for distinguishing one kind of group from another, in terms of pathology, is modest at best. As the manual states, "A high score on a scale has been found to predict positively the corresponding final clinical diagnosis or estimate in more than 60% of new psychiatric admissions." (Id.) It does not take a mathematical genius to figure out that the MMPI fails to accurately predict personality disorders in almost 40% of the cases. Should such predictive error be allowed in the courtroom?

There is no research indicating that the MMPI is 100% valid or totally invalid. The weight of the psychological literature probably could be summed up by saying the MMPI is a valuable clinical tool but should not be relied on to the exclusion of everything else.

Consider some of the descriptions quoted by Jay Ziskin in his work, Coping with Psychiatric and Psychological Testimony. Ziskin discusses the validity and reliability of the MMPI as follows:

> "The test is dependent for its power on self-description. It was empirically developed from patient populations that were reasonably cooperative and reasonably motivated to reveal upset. In a differently motivated population, the test and its standard norms are not valid and can be grossly misleading.
>
> Validity has always been held to be the most important aspect of a test review, yet in this case, it is a most difficult matter to evaluate. We might well decide that the MMPI as an instrument is valid for many interpretations and purposes but a varying levels of effectiveness. This would be my position in general. Yet it would also be possible to doubt the validity or "adequacy" of a global personality description based on the MMPI alone. Relatively few studies have made concerted attacks on the problem of global validity. These were encumbered by methodological problems, but, regardless of this, results were quite disappointing. Using the devices of Q sorts and true-false rating scales, test judges agreed with therapists and/or interview judges with no greater correlation than 0.40. Yet, it is with the aspect of "global" validity that we are most concerned. It is a complex question." Ziskin, supra, at 220-21. (Emphasis added.)

Although the MMPI is described as a "standardized, objective" test, the terms can be misleading. All that "standardized" means is that the test asks the same questions in the same order of every person taking the test. All "objective" means is that the person taking the test has only three responses: true, false, or cannot say.

Regarding the MMPI as a whole, the test may be a good clinical tool for diagnosing and treating mental health patients. However, the question arises as to whether the test is a good indicator of who is the best parent. There are few items dealing with "the family" on the MMPI and fewer still about children. There are no items that say, "I love my child" or "I'd rather see my son's baseball game than play golf."

While the new normative sample developed by Colligan, et al., is a probable improvement over the MMPI in that a new sample group of normals was established, what about the abnormals? The work did not create new psychiatric samples. It would seem, therefore, that criticism of the MMPI based on the scant information concerning some of the small samples of psychiatric patients done in the 1940's would still be valid. Additionally, many of the items seem outdated in their content.

Further, even though there is a new normative sample, the sample group of 1,408 people are all form Minnesota. Colligan reports that "with the possible exception of five subjects for whom we have no information, all our subjects were white." Colligan, 1983, page 83. Statistical theory notwithstanding, is it fair to compare the results of the testing of one Hispanic from Houston to the testimony of 1,408 whites from Minnesota?

One last general observation concerns diagnostic language and categories. Many of the abnormal groups used to create the ten clinical scales on the MMPI were diagnosed as having disorders which have outdated labels. Trying to find where the schizophrenic of 1940 fits into DSM III (Diagnostic and Statistical Manual of Mental Disorders, 3rd Ed.) is quite difficult. If the definitions have changed, shouldn't the clinical scales and their development also be updated? (McCurley and Fuller, supra, at T-48.)

4. MMPI-2

a. Differences from the Original MMPI.

According to Beverly Kaemmer, a publications representative of the University of Minnesota Press, not much has changed with the revisions, other than removing items that were thought to be "offensive" such as questions about religion, sexual practices, and bowel and bladder functions. In fact, the items that were deleted never scored on the basic scales. A small percentage of the items were revised to eliminate sexist language, make the questions more clear, and make the content more approachable. Other items have been designed "to assess such behavior as treatment compliance and amenability to change." (S. Hathaway and J.C. McKinley, "Fact Sheet on the Minnesota Multiphasic Personality Inventory - 2 (MMPI-2)", MMPI-2 WORKSHOP, Section 3, University of Minnesota, 1989.) One proposed revision includes 100 new items for a version of the exam geared specifically toward adolescents -- an important revision to note when considered in conjunction with an adolescent's ability to file a "Choice of Managing Conservator" affidavit with the Court.

In arriving at the changes, a research form of the MMPI booklet was developed, which contained not only the original 550 items, but an additional 154 items, most of which:

> ". . . were intended to replace ones that were culturally outmoded or psychometrically unsound in the existing inventory as well as to serve as sources of supplementary measures in the areas of family dynamics, Type A behavior, eating disorders, substance abuse, suicide and readiness for treatment rehabilitation." (Id.)

The experimental booklet, dubbed "AX" as an amalgam for "Adult Experimental" was conducted as a parallel study to the adolescent MMPI study referenced above. In addition to the experimental booklet(s), subjects were asked to complete special supplementary forms, which were designed to assess recent changes in the subject's lives, including the degree of satisfaction the subject had in his or her personal relationships. (Id.)

The sample group which took the new experimental versions of the MMPI ranged in age from 16 to 90; the normative sample group was whittled down to 1138 males and 1462 females. One problem that the committee in charge of the re-standardization still has arguably not surmounted is the fact that the "Hispanic - and Asian-American subgroups are under-represented in the re-standardization sample." (Id.)

The biggest change with the 1989 update to the MMPI is in the norms, or samples. Rather than limiting the norms to tests conducted in Minnesota, the norm is now based on a nation-wide sample. As to the nation-wide sample, Anne Anastasi writes that:

> "The two experimental forms of the MMPI were administered to nationally representative normative samples, including approximately 3,000 persons in the adult sample and 3,000 in the adolescent sample. The participants comprise random samples within communities chosen because their demographic characteristics conform closely to the 1980 U.S. Census. The total normative sample was drawn from several regions of the United States, including the states of Minnesota, Ohio, North Carolina, Pennsylvania, Washington, Virginia and California. The aim was to obtain a sample that is nationally representative according to urban-rural and geographic residence and such demographic characteristics as age, sex, educational level, and ethnic group membership; preliminary analyses indicate a fairly close match with the 1980 Census in these demographic variables." (A. Anastasi, Psychological Testing, 553 (6th Ed. 1988.)

The MMPI-2 and the MMPI are very much the same, but the following refinements distinguish the MMPI-2:

(1) Implementation of national norms.

(2) T scores are based on eight of the basic scales, refining Scales 5 and 0 (masculinity/femininity and social introversion), and generating uniform T scores to produce the same type of two-point and three-point high-point codes as have been used in the past.

(3) Although the test booklet is about the same length (567 questions), sexist language and objectionable items have been modified; the 16 duplicate items have been deleted; and the item order is changed, thereby facilitating a score of the basic scales based on the first 370 items. The remainder of the items provide supplementary material, and although some items were on the original exam, many augment the exam.

(4) More subtle indications of personality will be assessed by new scales intended to assess protocol validity, to provide separate measurements of masculine and feminine gender roles, and to provide diagnostic aids to respondents in clinical settings. (Hathaway, supra.)

Ms. Kaemmer, of the University of Minnesota Press, emphasizes that the changes are not expected to have that big of an impact, because they do not impact on the basic scales.

b. Construction: Modifications and Continuities - MMPI Versus MMPI-2.

Drs. Hathaway and McKinley boiled down the initial pool of 1000 items to 550 items or questions on the original MMPI. Sixteen of the items are repeated on the original MMPI as a fail safe device, for a total of 566 items. The items were subdivided into 26 different categories on the original MMPI, delineated below:

General Health (9 items)
General neurological (19 items)
Cranial Nerves (11 items)
Motility and Coordination (6 items)

Sensibility (5 items)
Vasomotor, trophic, speech, secretory (10 items)
Cardiorespiratory (5 items)
Gastrointestinal (11 items)
Genitourinary (5 items)
Habits (19 items)
Family and marital (26 items)
Occupational (18 items)
Education (12 items)
Sexual attitudes (13 items)
Religious attitudes (19 items)
Political attitudes - law and order (46 items)
Social attitudes (72 items)
Affect, depressive (32 items)
Affect, manic (24 items)
Obsessive, compulsive (15 items)
Delusions, (31 items)
Phobias (29 items)
Sadistic, masochistic (7 items)
Morale (33 items)
Masculinity - femininity (55 items)

Items to indicate whether the person is trying to paint himself as socially unacceptable (15 items) (M. McCurley and K. Fuller, supra.)

Once the items were differentiated into categories, the next issue to resolve was how to score the results -- was the test taker compatible with normal scores or those of the abnormal people?

The MMPI-2 contains 567 items, with no repeats.

5. Development of MMPI and MMPI Norms.

Guidelines for how to score results of the MMPI are based on three normative groups: the Hathaway and McKinley Group of 1940, the Hathaway and Briggs Group of 1957, and the contemporary normative study of Colligan, Osborne, Swenson and Offord, of 1983. (Id. at 7-9.)

(a) The Original Normative Group: Hathaway And McKinley, 1940 .

As indicated above, the initial sample was comprised of a specified clinical group and of a normal control group made up of visitors to the University of Minnesota Hospital and other groups from the Minnesota area, including persons attending pre-college conferences at the University of Minnesota, and WPA administration workers. The control group represented a cross section of both sexes of the Minnesota area between the ages of 16 and 55. (A. Anastasi, supra, at 527.)

(b) The 1957 Group: Hathaway And Briggs .

This normative grouping was comprised of a portion of the original 1940 group, and consisted of 226 males and 315 females. Presumed to be "normal," the range of the 1957 group's responses to the MMPI was derived from a comparison and contrast with the responses these two groups of people originally gave to the 1940 exam. [Hathaway and Briggs, "Some Normative Data on the New MMPI Scales," 364-368 Journal of Clinical Psychology, Vol. 13 (1957).]

For a thorough review of the difference between the original 1940 normative group and the 1957 normative subset, the authors recommend further reading in Colligan, Osborne, Swenson and Offord, The MMPI: A Contemporary Normative Study, 12-14, 68 (New York: Praeger Publishers, 1983).

(c) The Contemporary Normative Group: Colligan, Osborne, Swenson And Offord, 1983.

Colligan and his colleagues developed a new range of responses for the MMPI. The Colligan group also created a new manner of calculating T scores. In a nutshell, if a person scores 70 or higher on any scale, that person has responded differently from 97.7 of the "normal" subjects taking the MMPI. A score of 70 or above on any scale is therefore considered "clinically significant." [Colligan, Osborne, Swenson & Offord, supra, at 71-91.]

6. Creating the MMPI-2 Scales.

Originally, the MMPI scales were intended to differentiate between what was considered "normal" and what reflected traditional diagnostic categories. Just because your client happens to score high on the paranoia scale does not necessarily mean that he or she is paranoid. Therefore, an elevated score on the paranoia scale may simply indicate the test taker tends to be distrustful, investigative and curious. [J.T. Kunce and W.P. Anderson, "Perspectives and Uses of the MMPI in Non-psychiatric Settings" in P. McReynolds and C.J. Chelune (Eds), Advances in Psychological Assessment, Vol. 6, 41-76 (San Francisco: Jossey-Bass, 1984).] Further research has illustrated that a single elevated score taken out of context may not truly reflect the test taker's personality, and that a multidimensional, overlapping review of MMPI scales is the preferred approach. The re-standardization committee charged with compiling the MMPI-2 opted for such a multi-dimensional approach.

7. Other Scales.

Since the 1940's over 300 new scales have been created, most by independent investigators who were not privy to the development of the original MMPI. The new scales are varied. Many scales were developed by using normal samples to assess personality traits that were not related to pathological constraints. As Anne Anastasi notes:

". . . Some scales have subsequently been applied to the test records of the original MMPI normal standardization sample, thus providing normative data comparable to those of the initial dependency scales. Examples of the new scales include: Ego Strength (Es), Dependency (Dy), Dominance (Do), Prejudice (Pr), and Social Status (St). Other scales have been developed for highly specialized purposes and are more limited in their applicability. Still another grouping of MMPI items is represented by the content scales developed by J.S. Wiggins. In the construction of these scales, item clusters based on a subjective classification of content were revised and refined through factor analytic and internal-consistency procedures. The resulting 13 scales have proved promising in diagnosis and may serve as a useful supplement in the interpretation of the original scales." (A. Anastasi, supra, at 526.)

Frequently scored supplementary scales on the MMPI-2 include the following:

(a) ANXIETY (A)

Anxiety is a factor dimension that emerges when the clinical and validity scales are factor analyzed. High scores may indicate maladjustment, lack of social poise, or that the person is inhibited, cautious, distant and uninvolved; if male, high scorers tend to be effeminate, and may become confused and disorganized under stress. Low scorers are active, vigorous, well-spoken, resourceful, power-oriented, able to lead and even manipulate others, and prefer action to thought. [G.S. Welsh, "Factor Dimensions A and R." In G.S. Welsh and W.G. Dahlstrom (Eds), Basic Readings on the MMPI in Psychology and Medicine (Minneapolis: University of Minnesota Press, 1956).]

(b) REPRESSION ®)

Repression, along with anxiety, is a factor dimension that emerges when clinical and validity scales are factor analyzed. High scorers are considered submissive, internalizing, conventional and formal. Low scorers are outgoing, talkative, informal, jolly, impulsive and shrewd. (Id.)

(c) MANIFEST ANXIETY SCALE (MAS).

The Manifest Anxiety Scale was developed to identify subjects with high and low drive (i.e., anxiety) in order to study the effect of the level of the subject's drive on performance. High scorers are predisposed to experience emotional discomfort in stressful situations, jumpy, subject to excessive perspiration, emphasize the present more than the future, perform well on simple tasks and poorly on complex tasks. Low scorers tend to remain calm and unruffled in stressful situations, and are relatively free of physical and/or somatic complaints. [J.A. Taylor, "A Personality Scale of Manifest Anxiety," 48 Journal of Abnormal and Social Psychology 285-290 (1953).]

(d) EGO STRENGTH (Es).

The purpose of the ego strength supplementary scale is to assess the response of neurotic patients to individual psychotherapy. High scorers are tolerant, lack chronic psychopathology, have a secure sense of reality, and create a favorable first impression. Low scorers have poor psychological adjustment, are not well equipped to deal with stress, are more likely to be diagnosed psychotic than neurotic, and have chronic physical complaints and/or fatigue. [F. Barron, "An Ego Strength Scale Which Predicts Response to Psychotherapy" 17 Journal of Consulting Psychology 323-327 (1953).]

(e) DOMINANCE (Do)

The dominance supplementary scale seeks to identify people who are dominant in interpersonal relationships. High scorers are not readily intimidated, feel safe and secure, are task-oriented, and persevere. Low scorers are weak in face to face contacts, unassertive, pessimistic and inefficient. [H.G. Gough, H. McClosky and P.E. Meehl, "A Personality Scale for Dominance," 46 Journal of Abnormal and Social Psychology 360-366, (1951).]

(f) SOCIAL RESPONSIBILITY (Re)

The social responsibility scale evaluates willingness to accept responsibility for one's own behavior and sense of responsibility to the group. High scorers are willing to accept responsibility for the consequences of their behavior, trustworthy, have a deep concern for ethical and moral problems, reject privilege and favor, and have trust in the world in general. Low scorers are unwilling to accept responsibility for their own behavior, have flexible values, and lack dependability and trustworthiness. [Id. at 73-80.]

(g) COLLEGE MALADJUSTMENT (Mt)

This scale discriminates between emotionally adjusted college freshman and those students who are emotionally maladjusted, ineffectual, pessimistic and procrastinators. Low scorers are adjusted, conscientious and optimistic. [B. Kleinmuntz, "Identification of Maladjusted College Students," 7 Journal of Counseling Psychology, 209-11 (1960).]

(h) MACANDREW ALCOHOLISM (MAC).

This scale is designed to identify alcoholics and individuals prone to alcoholism. High scorers are more likely to be excessive in using drugs or alcohol, and may be prone to black-outs. Low scorers are less likely to use drugs, and are shy and conventional. [C. MacAndrew, "The Differentiation of Male Alcoholic Outpatients from Non-Alcoholic Psychiatric Outpatients by Means of the MMPI," 26 Quarterly Journal of Studies on Alcohol, 238-46 (1965).]

(i) OVER-CONTROLLED HOSTILITY (O-H).

This scale attempts to distinguish assaultive from non-assaultive prisoners. High scorers have strong emotional control, and report few angry feelings. Low scorers are emotionally immature, irresponsible, and have little emotional control. [E.I. Megargee, P.E. Cook & G.A.

Mendelssohn, "Development and Validation of an MMPI Scale of Assaultiveness in Over-controlled Individuals," 72 Journal of Abnormal Psychology, 519-528 (1967).]

8. Myths and Facts About Scoring The MMPI And The MMPI-2.

(1) MYTH: YOU CAN FAIL THE MMPI.
FACT: You cannot fail the MMPI.

A particular score on the MMPI does not mean that without question a person is sane, insane, normal or abnormal. In the worst case scenario in a custody dispute, a high or clinically significant score on the MMPI means that the test taker answered a question in a manner like that of a particular group of "abnormal" subjects or unlike that of a particular group of "normal" subjects.

Should a person score high on Scale 6 (Pa) for example, that person is not necessarily paranoid. Such a score would simply indicate that the test taker endorsed items similar to those endorsed by a statistically significant number of paranoids rather than those endorsed by a statistically significant number of normal subjects.

(2) MYTH: THE MMPI IS THE FUNCTIONAL EQUIVALENT OF A CLINIC.
FACT: The MMPI is a clinical tool.

Interpretation of the MMPI involves creating profiles from the T scores achieved by the individuals taking the test. The creators designed the MMPI such that scores would be achieved by comparing the interrelation of T scores of the 10 clinical scales, instead of by examining each scale individually. [Graham, The MMPI: A Practical Guide, 18 (New York: Oxford University Press, (1977).]

9. MMPI-2 Profile Scoring

MMPI profiles are created through the following steps:

(1) The raw scores are tabulated and converted into T-scores with "K" correction as appropriate. Remember, if the "K" scale indicates that the respondent was "faking," the T-score from the "K" scale is used to "correct" T-scores from the clinical scales.

(2) The T-scores are arranged in such a manner that the highest score appears first, the next highest second, et cetera.

(3) Depending on the sophistication of the clinician, "code types" are used to arrive at interpretations.

(4) An interpretation of "code-type" or interrelation of the scales is primarily premised on consulting prevailing psychological literature and clinical experience. As of 1977, there were over 9,000 books and articles published on the development, administration, and interpretation of the MMPI. (Id. at 64.)

The MMPI-2 re-standardization weighed two possible approaches to scoring. The first involved a manipulation of normalized T-scores, as advocated by Colligan, Osborne, Swenson and Offord. The re-standardization committee decided that the manipulation of standardized T-scores would alter the MMPI profile, and further that such a manipulation was not justified with the MMPI-2, as the normative sample used to judge everything was large. The re-standardization committee therefore opted for another approach, which involves:

> ". . . deriving a composite (or average) distribution of the raw scores on the eight basic clinical scales, and adjusting the distribution of each clinical scale so that it would match the composite distribution. The implementation of such a procedure resulted in a set of uniform T-scores that are percentile equivalent . . ." (Y.S. Ben-Porath, MMPI-2 Consistency Scales , MMPI-2 Workshop, § 3. Univ. of Minn. 1989.)

10. Computer Scoring of the MMPI and MMPI-2

The MMPI can now be scored by means of a computer. The process is relatively simple. The completed answer sheet of an examinee is fed into the computer, usually by optical scanning equipment. The protocol is scored, and the appropriate classification rules are applied in order to determine the categories for the profile. The computer then searches its memory for interpretive statements appropriate for the categories and prints out a report. [Graham, supra, at 185-86.]

Sounds easy as 1, 2, 3. It is, except for the lawyer who would like to figure out the basis of the clinical opinions that resulted in an unfavorable MMPI profile for his/her client or a favorable MMPI profile for the other side. The problem is nobody knows who is doing the "Interpretation" within the internal computations of the computer program which generates the final report (except the computer programmers, the computer interpretation services and their copyright lawyers). With so much research in the field, some conflicting, some consistent, some validated, some not validated, some reliable and some unreliable, how are we, as lawyers, to intelligently examine the basis of computer interpretation? Is it Alter's, Barron's Baughman's, Black's, Block's, Boerger's, Butto's, Byrne's, Calvin's, Carkhuff's, Carson's, Chu's, Comrey's, Cuadra's, Dahlstrom's, Davis', Distlers', Drake's, Duckworth's, Dunbar's, Edwards', Eichman's, etc. or some combination of the above? (The above is a partial list of research material cited as references as Graham. Graham, 1977, page 213-14.) (McCurley and Fuller, supra, at T-42.)

Graham lists six computer reporting services in his work and also described the various services offered by each of his entitles. Graham, 1977. It is also interesting to note that Graham sent the exact same MMPI test results to all six of the computer services. Although Graham reported general consistency among the results, there were some discrepancies. See Graham, 1977, pages 210-11.

One clear contradiction was observed by Graham in the report results. Four of the reporting companies stated that the subject was "Likely to report somatic complaints that have no clear organic basis. Graham, 177, page 210. One of the reports made no statements regarding

somatic complainings and another stated the subject is "not at all the sort of person that gets bodily symptoms to symbolize his emotional conflicts without organic cause." Graham, 1977, page 210 (emphasis added). In psychological jargon, this statement is contradictory to the other four reports.

In reference to the validity of computer generated interpretations of the MMPI, Ziskin quotes the opinion of one researcher as follows:

> "At this stage computerized narratives using psychological-test-based information is little more than an art (or craft) disguised as science. For the most part, the narrative reports are clinical hunches (often many steps removed data) which are automatically cranked out by an electronic beast that will, without conscience, weave a devastating and sometimes contradictory tale about an individual's personality and problems. The computer is a generally willing and efficient servant that will readily combine and give back scores of information from its vast memory. It cares not at all whether the information stored is from astrology charts, MMPI code books, Rorschach Indices, or Somatotype descriptors ... the "artisan" nature of this endeavor has been demonstrated, the "clinical" astuteness is often compelling, but the "science" is often neglected or of a tertiary consideration ..."
>
> "... By far the most haunting problem and serious shortcoming of the automated MMPI assessment approach remains that of system validation. Demonstrating the validity of computer-generated narratives (like that of demonstrating clinical interpretations generally) is a formidable task." Ziskin, 1981, page 223. (Emphasis added.)

(McCurley and Fuller, supra, at T-42.)

There are numerous MMPI computer reporting services. Perhaps the most common in use by clinical psychologists are: National Computer Systems; P. O. Box 1294; Minneapolis, Minnesota 55440; telephone, 612-933-2800; and the Caldwell Report Clinical Psychological Services, Inc.; P. O. Box 24624; Los Angeles, California 90024; telephone, 213-478-3433.

11. Advantages Of The MMPI And The MMPI-2.

The MMPI tests are easy to administer, and unlike many intelligence tests that require administration by the clinician, they can even be administered by a layperson in most circumstances. The MMPI and the MMPI-2 have long been recognized as excellent clinical tools for diagnosing and treating mental health patients. Moreover, the original MMPI is the most widely used personality inventory, and has the added benefit of years of empirical research being generated to help refine and interpret interpretation of the test. The test has the added benefit of the validity scales, constructed to catch attempts to deceive the examiner.

12. Disadvantages Of The MMPI And MMPI-2.

Implicit in the construction of most personality tests is the assumption that an individual's behavior is characterized by consistency, regardless of the situation. While intellectual facilities

are fairly consistent despite the situation, an individual's behavior may show considerable change from one situation to the next. "Thus, tests that identify individual traits without qualifying the situation in which they are likely to be manifest are probably misleading if not invalid." (Shuman, supra, at 62.)

Regarding the MMPI and MMPI-2 as a whole, the tests may be a good clinical tool for diagnosing and treating mental health patients. However, when you examine their validity in a custody dispute, their relevance must be questioned. As was previously stated, there are few items dealing with "the family" and fewer still about children. The MMPI and MMPI-2 **WERE NOT** DESIGNED TO HELP COURTS TO DETERMINE WHO IS THE BETTER PARENT!

The original MMPI has been heavily criticized for its sample group of norms and because some of the items seemed outdated in their content. (M. McCurley and M. McCurley, Psychological Testing Marriage Dissolution Course, N65-66 May, 1987.) The sample group of norms for the MMPI-2 appears to solve some of the problems that have plagued the MMPI. The new sample group consisted of larger segments of the population of the United States, crossing cultural barriers. The items in MMPI-2 have remained fairly untouched so the most of the criticism of the original items remain unchanged.

The last general observation concerns diagnostic language and categories. Many of the abnormal groups used to create the original ten clinical scales were diagnosed as having disorders which now have outdated labels. The clinical scales on MMPI-2 have remained virtually the same, as in MMPI. Trying to use DSM III-R (Diagnostic and Statistical Manual of Mental Disorders) in reviewing the clinical scales is quite difficult. If the definitions have changed, shouldn't the clinical scales and their development also be updated? (M. McCurley and K. Fuller, supra.)

13. Cross-examination of the MMPI and MMPI-2 Expert

Some suggested lines of questioning are as follows:

(1) Was the MMPI in question scored using the updated norms established by Colligan, et al., in 1983?
   - (a) If yes, is there research validating the new norms?
   - (b) If no, point out criticisms of old norms appearing through Colligan, 1983, and Ziskin, 1981.
   - ©) If the test was scored by computer, does the computer service incorporate the results of the new norms established by Colligan, 1983?

(2) Was MMPI scored, whether by computer or by the clinician, using the new calculations regarding T-scores developed by Colligan in 1983?

(3) How much does the expert rely on the MMPI in the formulation of his/her opinion?
   - (a) If the answer is 100% or "very much," the expert is in trouble and should be confronted with the research cited in this article.
   - (b) If the answer is only partly or only as one aspect, the expert is likely making proper use of MMPI results.

(4) Establish which MMPI was administered (short form, long form, etc.)

(5) Establish how many and what scales were scored.

(6) Attempt to get the expert to admit the limits of the reliability of the test. If the expert will not concede the limitations, confront the expert with the Green table cited herein. Also, note Ziskin states, "Reliability coefficients for the MMPI generally cluster in the area of the 0.70's which is low." Ziskin, 1981, page 217.

(7) If the expert is familiar with the lie detector test and its reliability and validity, have the expert give an opinion as to which is a more reliable and valid test as far as prediction is concerned. If the opinion is that the lie detector is more reliable and valid, why should MMPI results be admissible and lie detectors not. (See Ziskin, 1981, page 4, for such a comparison.)

(8) Subpoena the test administered and the party's answers. Questions such as, "So, you're saying just because Mrs. Jones says she doesn't have a satisfactory sex life, that means she is a psychopathic deviant?" should be easily fielded by a competent psychologist. However, for purposes of cross-examination of the opposing party, it might be interesting to know and point out answers to such items as "Evil spirits possess me at times."

Additionally, 16 of the MMPI items are repeats. Those items are:

1. 8 - 183;
2. 13 - 290;
3. 15 - 314;
4. 16 - 315;
5. 20 - 310;
6. 21 - 308;
7. 22 - 326;
8. 23 - 288;
9. 24 - 333;
10. 32 - 328;
11. 33 - 323;
12. 35 - 331;
13. 37 - 302;
14. 38 - 311;
15. 305 - 366; and
16. 317 - 362.

Collingan, 1983, page 67. (It should be noted that the repetition of these items was done so the MMPI could be more economically scored by the IBM 805 machine, not because of intentional design. Colligan, 1983, page 67.) If the answers to the repeat questions are not consistent, regardless of the scores on the validity scales, a good argument could be made against test validity.

(9) If the MMPI was computer scored, ask if the expert knows the basis of the interpretation placed on the scores by the computer; i.e., what research does the computer endorse? Dahlstrom's? Green's? Graham's? Hathaway's? etc.

(10) Was the test administered in the expert's office or was the party allowed to take the test home. If the test was taken at home, how does the expert know who took the test? (It would be rare that a home test would be permitted; although, note that in establishing the new norms, Colligan mailed the MMPI's to the subjects who took the test at home. Colligan, 1983, pages 75 and 333. This may be fruit for cross-examination itself.)

(11) If sub-scales were used, were the test scores different on the sub-scales than the standard clinical scales? Remember the sub-scales are "content" oriented instead of "empirically" developed.

(12) If the expert diagnoses your client as a manic-depressive (or whatever), determine what the expert observed or was told about your client's behavior and actions which supports this diagnosis independent of the MMPI scores. Attack the opinion with evidence of opposite behavior if it exists.

(13) Other Suggestions. As noted by the research done by Green, MMPI scores may vary over time. In a case where your client has fared poorly on the MMPI, a retest done by a consulting psychologist may look substantially different and provide good information for cross-examination. Note the validating scales on the MMPI should pick up any deliberate attempt to make the test results look different.

In a case where MMPI results are pivotal, sending the MMPI to the various computer scoring services may (small chance) result in different interpretations of the same test. The same could be done by taking the test results of another clinician and getting a second opinion. (McCurley and Fuller, supra at T-52.)

A valuable source for statistical information and reviews of MMPI test validity and reliability is Buros, O. K., supra. For valuable research information concerning the reliability and validity of MMPI short form tests, one should consult Green, R. L., Some Reflections on MMPI Short Forms: A literature Review, Journal of Personality Assessment, Vol. 46, No. 5, 1982, pages 486 - 487. (Green basically opines the imperfections of the standard MMPI are only magnified by use of short forms.)

B. 16PF

1. History, Background & Purposes of 16 PF

Through the use of a statistical process called "Factor Analysis," the 16 Personality Factor Questionnaire (16PF) has been developed to assess personality traits. Marc J. Ackerman, Ph.D. & Andrew W. Kane, Ph.D., How To Examine Psychological Experts in Divorce and Other Civil Actions 200 (1990). It is designed for use with individuals that are sixteen years of age or older and to analyze such traits as "reserved versus outgoing," "humble versus assertive," and "trusting versus suspicious." Id. These scales are short, the information is on normative samples, and the test construction is inadequate. Id. The test never gained the acclaim that the authors had hoped. Id.; see also, A. Anastasi, supra, at 542-43.

R. B. Cattell, in an effort to arrive at a comprehensive description of personality, began assembling all personality trait names occurring either in the dictionary, as compiled by Odberg (1936), or in the psychiatric and psychological literature. See Anastasi, supra. Cattell developed 171 trait lists which he employed first on a heterogenous group of 100 adults. Thereafter, he used intercorrelations and factor analysis, and conducted further tests on 208 men on a shortened list. From these factorial analyses, Cattell was able to identify what he described as "the primary source traits of personalties." Id.

Cattell's findings have been criticized as not true identifications of personality traits, but instead as reflections of the influence of social stereotypes and other "constant errors of judgment." Id. In fact, some scientists have found the same factors when analyzing ratings given to complete strangers as when analyzing ratings assigned to persons whom the raters knew well. (Pessini & Norman, 1966). Id. It is arguable that factor analysis of ratings may reveal more about the raters than about the ratees. Id.

Cattell, on the basis of his factorial research, has constructed a number of personality inventories, of which the best known is the 16 Personality Factor Questionnaire (Cattell, Eber, & Tatsuoka, 1970). Id. Cattell's test was designed for ages 16 and over, and yields 16 scores in such traits as "reserve versus outgoing," "humble versus assertive," "shy versus venturesome" and "trusting versus suspicious." Id. A "motivational distortion" or verification key is also provided for some of the forms. A computerized, narrative reporting service is also available for users.

Similar inventories have been developed for ages 12 to 18 (High School Personality Questionnaire), 8 to 12 (Children's Personality Questionnaire), and 6 to 8 (Early School Personality Questionnaire). Id. Separate inventories have also been published within more limited areas, including anxiety, depression, and neuroticism. Id. These areas correspond to certain second-order factors identified among correlated first-order factors. Id.

Another addition to the series is the Clinical Analysis Questionnaire, a 28-scale inventory which includes: a shortened version of the 16PF "in clinical dress", with fewer items per factor, reworded to fit clinical context; and 12 pathological scales identified through factor analysis of items from the MMPI and other clinical scales. Other adaptations of the 16PF in special dress have been prepared for assessment purpose in such context as career development, marriage counseling, and the evaluation of business executives. Id. All of these inventories are experimental instruments requiring further development, standardization, and validation.

The assessment of response bias is especially important in forensic psychological assessments and constitutes one of the unique advantages of psychological testing in this field. This is because in forensic evaluations, by definition, subjects generally have substantial, tangible gains, from either accentuating their strengths or their weaknesses. For this reason, psychological tests constructed with validity scales designed to detect response bias are particularly useful.

2. Reliability and Validity of the l6PF

Reliabilities of factor scores for any single form of the 16PF are generally low. Id. at 543. Parallel form reliabilities center around .50 and retests after a week or less often fall below .80. Id. Others have questioned whether the factorial homogeneity of items within each scale is reliable. (Lovenian, 1961). Id. Available information on normative samples and other aspects of test construction is inadequate. Id. Empirical validation data include average profiles for more than 50 occupational groups and about the same number of psychiatric syndromes. Id. "Specification equations" are provided for a number of occupations, in the form of multiple regression equations for predicting an individual's criterion performance from scores on the 16PF. Id.

In general, validity scales are designed to assess types of exaggeration or minimization in actual clinical evaluations. However, most studies of the validity scales of the 16PF have used experimental samples, such as students instructed to "fake good" or "fake bad" while taking the test, (See e.g., Braun & LaFaro, 1969; Riggio, Salinas, & Tucker, 1988; Stricker, 1974; Winder O'Dell & Karson, 1975) or job applicants undergoing psychological screenings (Birenbaum, 1986; Birenbaum & Montag, 1989; Elliott, 1976A, 1976B; Kochkin, 1987). Only a few studies exist on the use of the 16PF with forensic samples. (Audubon & Kerwin, 1982; Dalby, 1988; Irvine & Gendreau, 1974).

Because of the scarcity of research on the efficacy of the 16PF's validity scales in forensic settings, little is known about their efficacy in detecting clinical patients who attempt to "fake good" or "fake bad" (Green, 1988). For example, there have been few studies (Dalby, 1988) that have assessed the effectiveness of the 16PF in detecting response bias in forensic patients who are known to minimize or exaggerate psychopathology on the Minnesota Multiphasic Personality Inventory (MMPI) (Hathaway & McKinley, 1967), which has the most widely researched and effective validity scales of any psychometric instrument (Greene, 1980, 1988; Ziskin & Faust, 1988).

In addition, the 16PF literature has not produced a consensus about the best cut-off scores to use in deciding whether a particular profile shows significant response bias (Krug, 1978; Winder et al, 1975). The original cut-off score proposed by Winder et al (1975) was set at seven or higher (out of a possible 15 items per scale) for both motivation distortion (fake-good) and the fake bad scale. Krug (1978), using a larger normative sample for cross-validation, later suggested that a score of ten constituted a better cut-off criterion for the fake-good scale because this score classified 15% of his normative sample as a attempting to fake-good when taking the 16PF under standard instructions. Use of this rationale, however, assumes a normal distribution of response bias in the population being tested. That assumption is not a valid one in the population of parents being evaluated for custody.

As Grossman, Haywood & Wasyliw (1992) concluded in their study of the use of 16PF validity scales in the forensic psychological evaluations of alleged sex offenders, validity scales of the 16PF were significantly correlated with those of the MMPI and correctly predicted a high percentage of patients who showed minimization and exaggeration on the MMPI. Linda S. Grossman, Thomas W. Haywood & Orest E. Wasyliw, *The Evaluation of Truthfulness in Alleged Sex Offenders Self Reports: 16PF and MMPI Validity Scales*, JOURNAL OF PERSONALITY

ASSESSMENT 273 (1992). Their study further concluded that minimization is far more common among sex offenders than exaggeration in comparison to normal populations. Id.

In conclusion, factor analysis provides a technique for grouping personality inventory items into relatively homogeneous and independent clusters. Such a grouping should facilitate the investigation of validity against empirical criteria and such contribute toward construct definition and permit a more effective combination of scores for the prediction of specific criteria. Homogeneity and factorial purity are desirable goals in test construction, but are not substitutes for empirical validation.

B. Projective Personality Tests

Unlike the "objective" personality tests, projective tests present the testee with a stimulus and ask for a response. These tests involve more use of the psychologists's judgment than an "objective" test. These type of tests are generally more stressful to take because they are less structured than, for example, the MMPI. (Shuman, supra, at 51.)

R. B. Stuart, in describing projective tests, author of Trick or Treatment, states, "...the clinician who would use projective tests must answer Meehl's (1954) question, 'Am I doing better than I could by flipping pennies?' An honest answer to this question must be a qualified 'No' (page 89)." [R. B. Stuart, Foreword to Trick or Treatment, (1970).] It is with the above thought in mind, that this article explores the disadvantages of projective tests.

1. Rorschach Inkblot Technique

a. History and Background of the Rorschach

The most widely used projective test is the Rorschach. (Ziskin, supra, at 226.) The Rorschach was first introduced more than 50 years ago by a Swiss psychiatrist named Hermann Rorschach.

The test consists of ten inkblots printed on separate cards. Five of the cards are colored and the rest are black and white or shades thereof. Rorschach selected these particular cards after extensive clinical research, which led him to believe that this method of testing could help in personality study. Rorschach's results were based on test data from more than 400 subjects. (Garfield, supra, at 170.)

b. Construction, Administration, and Scoring of the Rorschach

The test consists of two main parts: (1) the free association period, and (2) the "inquiry." In the first part of the test, each of the ten cards are numbered and are shown one at a time in sequential order. The subject is asked to tell the clinician what he or she sees. All responses, as well as the time taken to respond to each card, are recorded by the examiner. After the free association period is over, the second part of the test, the "inquiry," is begun. The subject is shown the cards once again, but this time is asked by the clinician which parts of the card he responded to and what led him or her to perceive the blots as he or she did.

The "inquiry" part of the test is used to ascertain two things, "location" (what part of the blot was used), and "determination" (whether the form is judged good or bad, whether there is movement in the response, whether other features were used, i.e., color, shading, texture, etc.) of each response. These two factors, plus some designation of the actual content of the response, i.e., whether the subject sees an object, animal, or human being, constitute the minimum scoring of one response. A fourth scoring element is sometimes added when a response is considered to be popular (P) or often used. Various symbols are used to score these four elements of the test. These combined responses are then combined into what is called a "psychogram," which is a summary of all the scored responses. (Id. at 174.)

The scoring of this test is obviously not an easy matter. The test is, in great part, dependent on the abilities of the individual examiner. There are several scoring manuals available, but the development of the Exner system of administration and scoring in 1974 has been considered to have made a substantial improvement in the reliability and validity of the Rorschach technique.

Exner collected normative data from normal individuals and various diagnostic groups from a wide range of age groups (5 - adult). Because of Exner's system, a clinician can compare the individual who has been tested to normal people as well as to those with mental disorders. (Shuman, supra, at 52.)

c. Reliability & Validity of the Rorschach

Although the Rorschach is frequently criticized because of problems with reliability and validity, the Rorschach continues to be the most widely used projective test. Despite the popularity of the Rorschach with clinicians, Arthur R. Jensen, professor of Educational Psychology and Research Psychology at the University of California, Berkeley, has written the following review of Rorschach reliability:

> "Put frankly, the consensus of qualified judgment is that the Rorschach is a very poor test and has no practical worth for any of the purposes for which it is recommended by its devotees... There are now a number of methodologically and statistically sound and sophisticated studies. Even more important in terms of doing full justice to the Rorschach is that the good research is now being done by the Rorschachers and projective test experts themselves, often with the full cooperation of their clinical colleagues who are highly experienced in the use of projective techniques. No longer can it be claimed that negative findings are the result of bluenose methodologists of statistics and experimental psychology, applying inappropriate criteria to an instrument for which they have no sympathy nor clinical experiences, nor intuitive feeling, and no talent. (page 501)
>
> "In addition, the would-be Rorschacher, if he is to hold his own among the experts, must possess the kind of gifts similar to the literary talent of a novelist or biographer, combining a perceptive and intuitive sensitivity to human qualities and the power to express these perceptions in subtle, varied, and complex ways. The Rorschach report of an expert is, if nothing else, a literary work of art. This is the chief criterion of expertness with the Rorschach, for the research has not revealed any significant differences in reliability or validity between beginners in the Rorschach technique and acknowledged masters. (page 502)
>
> "Few other tests provide so many opportunities for the multiplication of error variance as does the Rorschach... In the typical protocol, most of the scoring categories are used relatively infrequently so their reliability is practically indeterminant... Most of the combinational scores from the Rorschach consisting of the ratios and differences among the various primary scores (page 504). Scoring reliability per se has been determined very seldom. Reliability of scoring

depends, to a large extent, upon the degree of similarity of the "training of the scorers and has been reported as ranging from 0.64 to 0.91... The most extensive determination of retest reliability is that of Epstein and others who gave he Rorschach to 16 college students, a total of 10 times over a period of 5 weeks. The average reliabilities for various response categories ranged from 0.29 to 0.56... Examiner and situational influences have been increasingly recognized in recent research as significant contributors to the variance of the Rorschach scores. The subject-examiner interaction is certainly one of the most important aspects of the test. The effect of the setting in which the test is taken and the fact that different examiners consistently elicit different amounts of various score determinants from subjects, should make it imperative that future Rorschach studies be based upon a representative sampling of examiners as well as of subjects...

"Reliability of interpretation is, of course, the most important matter of all. It may be stated as a general principle that the most crucial reliability is that of the end product of the test which, in the case of the Rorschach, usually consists of a verbal description of personality characteristics based on a global evaluation of all aspects of the subject's protocol. Contrary to the usual claim of Rorschachers that this global interpretation is more reliable or more valid than any of the elements upon which it is based, such as the scores and various derived combinations and indices, a systematic search of the literature has not turned up a single instance where the overall interpretation was more reliable than the separate elements entering into it. rorschach text books have not presented any evidence of satisfactory reliability of the final product of the test and the reviewer has not been able to find any such evidence in the research literature... Here are some typical examples of what has been found. Blasinsksky had 6 highly qualified Rorschachers rate 40 subjects on 10 personality items which they agreed could be confidently assessed from the Rorschach protocol. The correlation between the judges was 0.33. Six other clinicians rating the same traits on the basis of the case story abstracts alone, had a correlation of 0.31. The interesting point is that the 10 rated personality items were specially selected as being the kind of questions which the Rorschach, and not particularly the case history, is supposed to be able to answer." (page 505) (Arthur R. Jensen, In O. K. Buros,Sixth Mental Measurements Yearbook, 501 - 505.)

Jensen cites other studies of the Rorschach which show a mean co-efficient of 0.30, which is, of course, quite low.

It is important for the attorney to be aware of a study by Exner that some psychologists may cite as evidence of reliability of the Rorschach. The study appeared in the Journal of Personality Assessment in 1978 and was entitled, "The Temporal Stability of Some Rorschach Features." In Exner's study, 100 normal non-patients were retested after a three-year interval. Of 19 variables evaluated in this study, only seven reached the minimum correlation of 0.80. Therefore, should a clinician cite this study as proving the reliability of the Rorschach, he or she should be cross-examined regarding the fact that the majority of measures do not meet the

criteria for reliability. It should also be noted that even Exner acknowledges that there are a number of additional measures which were not even evaluated. (Ziskin, supra, at 241.)

Concerning validity, A. Anastasi sums up, for the attorney, what the research has shown is the validity of the Rorschach:

> "Nor can any encouragement be found in empirical studies of Rorschach validity. Despite a bibliography of over 2,000 publications on the Rorschach, the vast majority of interpretive relationships that form the basis of Rorschach scoring have never been empirically validated. The number of published studies that have failed to demonstrate a significant relation between Rorschach scores, combination of scores, or global evaluations and relevant criteria, is truly impressive. The Rorschach was found to have little or no predictive or concurrent validity when checked against such criteria as psychiatric diagnosis, response to psychotherapy, various determinations of personality or intellectual traits in normal persons, success or failure in a wide variety of occupations in which personality qualities play an important part, and presence of various conflicts, fears, attitudes, or fantasies independently identified in patients. Those studies that appear to provide positive results have been shown to contain serious methodological defects." [A. Anastasi, supra.]

Jensen points out that the validity studies of the Rorschach, which show it to be statistically significant, fall in the general range of 0.2 - 0.4 and are so low that one cannot conclude that the test has clinical usefulness. (Ziskin, supra, at 230.)

Jensen even suggests that, until proponents of the Rorschach can substantiate their claims of validity, the Rorschach should be abandoned in clinical practice and that students of clinical psychology "not be required to waste their time learning the technique." (Id. at 230.)

It appears obvious that surely, in more than fifty years of research on the Rorschach, some positive research regarding its validity should have been produced by now, and yet is seems there has not been. The question, then, of why the Rorschach has so many followers is a mystery.

Rolf A. Peterson, Professor of Psychology at the University of Illinois, asserts that it is not likely that Rorschach devotees will discontinue its use; it is, therefore, incumbent on them to provide the predictive validity data upon which their interpretations and predictions are based, for without such data, their conclusions are merely theory. (Id. at 235-36.)

d. Advantages of the Rorschach

Although developed to analyze basic personality structure, the Rorschach Inkblot technique has also been used by the psychological and psychiatric community to examine related aspects of personality such as attitudes, motives, aspirations, and conflicts. The value of the Rorschach as an aid in diagnosis is highly controversial. However, its usefulness may be, as Garfield puts it, dependent on "one's philosophy of diagnosis." (Garfield, supra, at 176.)

Garfield, Exner, and many others believe that the Rorschach has value at the clinical level because often the severity of a patient's problem is not apparent on initial interviews. The results of the Rorschach may provide information for a more complete evaluation of the patient.

e. Cross-Examination of the Rorschach Expert

Areas of cross-examination, other than the obvious questions regarding the reliability and validity of the Rorschach, are the following:

(1) Whatever the clinician proposes as his or her interpretation of the subject's responses to the test, he or she should be asked whether there is contradictory psychological literature or studies that challenge his or her assessment. If the clinician denies such studies, then the attorney must be armed with several studies that contradict his or her hypothesis. (The research material in this article is good background for many of such studies.)

(2) There are at least five major Rorschach methods or systems in the United States alone. The five systems - Beck, Hertz, Klopfer, Piotrowsky and Rapaport-Schafer - differ from each other enormously. They differ in basic administrative procedures, scoring, and interpretive hypothesis. Exner reports that a survey showed that of the five systems, roughly 54% of the psychologists prefer the Klopfer System, and 34% prefer the Beck System. (Id. at 237.) On cross-examination, the clinician should be asked which of the five major systems he or she used. If the clinician used the Klopfer System, it must be emphasized that almost half of all clinicians that use the Rorschach do not use that system. In the case of all other systems, less than half use any one system.

(3) Exner discovered that when a battery of tests is given, the sequence in which the tests are given can make a difference in the test results. The administration of one test affects the response given in different tests that follow. Exner cites studies that show, for example, that the incidence of human content responses on the Rorschach can alter depending on the order in which the different tests are given. Thus, the attorney should ask the expert witness if the results he or she obtained might have been different if given in a different order. (This is also true of other tests.) (Id. at 238.)

(4) It is highly questionable whether the adversary system is conducive to conjecture. Because child custody issues involve questions of a predictive nature, it is even more questionable whether we should allow conjecture in the courtroom. According to Exner, child custody questions are almost impossible to answer from most assessment data and particularly from Rorschach data. (Id. at 240.) The attorney should elicit testimony from the clinician who administered the Rorschach if the clinician recognizes Exner as a known expert in the field of psychology. The attorney should then ask the clinician if he or she knows that Exner believes that the rorschach is or little or no use in child custody disputes.

It is clear that the Rorschach is a highly controversial test, both as to the method of using it and even whether it should be used at all. Given the wide disagreement within the profession of psychology itself, "...no judge or jury should be asked to believe any conclusions based upon its use." (Id. at 242.)

2. Thematic Apperception Test (Tat)

a. History and Background of the TAT

The TAT was developed in the 1930's (Murray, 1943) and continues to be a popular technique used in clinical assessment. The TAT has been ranked as the fourth most frequently used psychological test. (B. Ritzler, K. Sharkey and J. Chudy, "A Comprehensive Projective Alternative to the TAT," Journal of Personality Assessment, 1980, at 358.)

The purpose of the procedure, when created, was to "evoke fantasies that reveal covert and unconscious complexes." (Id. at 530.) The test was based on the idea that when faced with an ambiguous social situation, a person will expose his or her own personality. The theory, taken further, is that once the per so n is explaining the occurrence, he or she becomes conscious of himself or herself, and therefore, becomes vulnerable to scrutiny.

The creators of the TAT came up with a set of pictures, that, in most instances, there was at least one person in the picture with whom the subject could, hopefully, relate. Therefore, there were separate sets of pictures chosen for children, males, females, young adults and older adults.

b. Construction of the TAT

Standard TAT instructions stress imagination and creativity. The test consists of 30 cards which depict people in different situations. A few of the cards show other types of scenes without people in them. The examiner may present the standard cards or may, selectively, choose his own if he or she has a particular area on which they wish to focus. The subject is then asked to look at each card and make up a story about it. (Shuman, supra, at 53.)

In the original study done by H. A. Murray, it was discovered that there were four main sources from which the stores were drawn: (1) books and movies; (2) actual events of which the subjects were aware; (3) experiences in the subject's own life; and (4) the subject's conscious and unconscious fantasies. (H. Murray, Explorations in Personality (2d ed. 1947) at 533.)

The first two sources, Murray felt, were those things which had the deepest impression on the patient and, therefore, comprised a clue to the subject's personality. As to actual events or experiences, Murray pointed out that "every subject almost immediately projects his own circumstances, experiences, or preoccupations." For instance, in one of the early experiments, six of the eleven college men who took the test said that the youth in one picture was a student; whereas none of the twelve non-college men who acted as subjects described him as such." (Id. at 533.) To analyze the subjects' possible fantasies, each story is read and diagnosed separately, and then an attempt is made to find a unifying theme.

c. Variations of the TAT

Most of the pictures in the TAT are in dark shadowy tones and most of the scenes are low-keyed and depressing situations. For example, in one picture, a short elderly woman stands with her back turned to a tall young man. The latter is looking downward and with a perplexed

expression, his hat in his hands. Or, in another picture, there is a heavily built man, naked to the waist with his head hung downward and his arms hanging limply at his side. Some researchers feel such somber pictures may elicit the negative thoughts of the patient.

Therefore, Barry Ritzler, Kevin Sharkey and James Chudy put together a new set of pictures; however, they used the same techniques of testing as used in the original TAT. The researchers used pictures taken from the Family of Man photo essay collection published by the Museum of Modern Art (1955). The following criteria was used to select each picture:

(1) The pictures had to show "potential for eliciting meaningful projective material" (Murray's only criterion).
(2) Most of the pictures had to include more than one person.
(3) At least one-half of the pictures had to show positive expression (e.g., smiling, dancing, embracing, etc.)
(4) At least one-half of the pictures had to show activity other than merely sitting or standing. (Ritzler, Sharkey and Chudy, supra, at 359.)

The results of the test illustrated that the new technique provided a balance of positively and negatively-toned stories. The researchers felt that they had developed a TAT which covers a more comprehensive range of human functioning and, therefore, they may make a fairer assessment for studying personality characteristics. (Id. at 361-62.)

d. Reliability and Validity of the TAT

The fundamental problem with the TAT is that it is subject to an even wider range of distorting factors than objective tests, or even of some other projective tests. For example, examiner bias, particularly in cross-sex administrations affect the type of content elicited. (Id. at 242-43.)

Leonard Eron, Professor of Psychology, University of Illinois, among others, states that the TAT is not suitable for providing a profile of personality traits or a reliable measurement of any one trait. Eron then states, "Nothing has appeared in the literature in the last five years which would serve to refute these conclusions with any degree of conviction." (Id. at 243.)

It is extremely important for the attorney to know, when confronted with a clinician with negative findings from the TAT about the client, that the TAT is still being published with the original manual (1943), which gives no reliability or validity data. (Id. at 243.) Some researchers have suggested that if the TAT were published today, with no information on reliability, validity and standardization, it would most likely not attain anywhere near its present popularity. (Id. at 243.)

J. D. Swartz, who reviews the TAT in the Eighth Mental Measurements Yearbook, explains that the body of TAT research does not provide any cohesive knowledge regarding the applications of the test to personality evaluation. Reliability and validity are especially a problem with the TAT because "...of the free nature of the response, the high degree of interaction with situational factors, great difficulty in attaining suitable independent criteria to

validate inferences against and problems of trying to develop meaningful quantitative measures." (Id. at 244.) Apparently, no one has been able to convince the majority of experts in this field that the above-mentioned problems with the TAT have been solved, which raises questions like those on the Rorschach about the usefulness of a test that has mixed research results after over forty years of investigation. Additionally, because of the many variations in scoring and administration, the TAT is likely to be as ambiguous to the examiner as it is to the subjects responding to it. Thus, the TAT, like the Rorschach, remains a test of controversy, with strong doubts as to its reliability and validity and its ability to assess individual personality traits accurately. (Id. at 244.)

e. Advantages of the TAT

The TAT can elicit information which is useful to diagnosis and treatment planing when it is used in conjunction with other test data and the history of the patient. Because of the recent research data regarding the negatively-toned stories resulting from the selection of the pictures by Murray, if the clinician uses the techniques by Ritzler, Sharkey and Chudy, then the result is an even more valid comprehensive assessment tool for the clinician.

f. Cross-Examination of the Expert

The line of cross-examination of the TAT should, as with the Rorschach, be focused on the reliability of the test. The attorney should elicit testimony which shows that anyone who administers this test in a child custody situation has done so when the great weight of authority shows the TAT results to be of little benefit in such situations and, at very best, full of conjecture.

Additionally, because of the concern among clinicians regarding the pictures that compose the TAT, the clinician should be questioned as to whether he or she took the studies regarding the negatively toned pictures into account when interpreting the data derived from the TAT. (Shuman, supra, at 53.)

3. Sentence Completion Tests

a. History, Background and Purposes of Sentence Completion Tests

Another verbal projective technique, sentence completion, has been widely employed in both research and clinical practice. Generally, the opening words for a sentence stem, permit an almost unlimited variety of possible completions. Examples might be: "My ambition_____; Women_____; What worries me_____; My mother_____." These sentence stems are frequently formulated so as to elicit responses relevant to the personality domain under investigation. The flexibility of the sentence completion technique represents one of its advantages for clinical and research purposes. Nevertheless, some standardized forms have been published for more general applications.

The use of the sentence completion technique as a psychological test is considered to have its origin with the works of Ebbinghaus (1897) in his studies of mental abilities. Binet and Simon (1905) found the method to be useful for measuring intellectual abilities and included sentence stems as one of the tests in their first battery. However, the sentence completion method is now primarily thought of as a technique for assessing personality and attitudes. For use in personality assessment, the test can be traced to the word association technique, which originated with the work of Jung in 1904 and 1906. While valuable material can be gathered through the word association technique, it was soon realized that limitations exist to this method. Longer and possibly more structured stimuli and responses are useful for the investigation of personality.

The first use of a sentence completion test for personality assessment is attributed to psychologist A. F. Payne in 1928. Others began to use similar instruments, but not until World War II did the sentence completion method become popular. A need existed for a quick, easily administered psychological test that could be given to large groups of people. The sentence completion method met the needs and soon became a part of psychological batteries in military settings. It was used in the Air Force as a screening device, but was probably best known for its use in the Office of Strategic Services. In nearly all settings, a sentence completion test was part of a battery of tests or part of a larger set of data about a person and often served as an aid in subsequent interviews. In military hospitals, it was used as a screen device to help decide who should be given more thorough psychological testing.

One such test, developed to be used in Air Force hospitals, was adopted for civilian use after the War. The Rotter Incomplete Sentences Blank (RISB) is one of the most widely used sentence completion tests in part because it was one of the few which allows both qualitative and quantitative assessment of responses.

b. Types of Sentence Completion Tests

(1) Rotter Incomplete Sentences Blank Test (Risb)

The Rotter Incomplete Sentences Blank consists of 40 sentence stems. See Anastasi, supra, at 608. The directions to the test-taker read:

"Complete these sentences to express your real feelings. Try to do every one. Be sure to make a complete sentence."

Each completion is rated on a seven point scale according to the degree of adjustment or maladjustment indicated. Id. With the aid of these specimen responses, fairly objective scoring is possible. The sum of the individual ratings provide a total adjustment score that can be used for screening purposes. The response content can also be examined clinically for more specific diagnostic clues.

Rotter intended to use the Incomplete Sentences Blank as an alternative to a lengthy structured interview. While the test was not seen as exposing extremely deep levels of personality, it was anticipated that information not readily available during an interview would often be brought out. By placing some distance between the examiner and the examinee via use of the test, it allows the examinee to respond more freely than he or she might in a face-to-face interview. The examiner can take the information at face value, yielding a quick overview of some of the issues for a given individual.

(a) Scoring

The scoring systems for all three forms of the test are based upon the scaled responses of college freshmen in the development of the College Form of the RISB. Each of the 40 items receives an "adjustment score." Each item is assigned one of four possible codes: (1) Omission (no response or too short a response to be meaningful); (2) Conflict response (reflecting hostility or unhappiness); (3) Positive response (reflecting positive or hopeful attitudes); or (4) Neutral response (no significant positive or negative effect). Both "conflict" and "positive" responses are given a weighing of one to three to reflect the degree of sentiment expressed. The manual provides manuals for each scoring category, with men's and women's responses listed separately.

(b) Advantages of the Test

This ease of administration and initial interpretation are prime reasons for the popularity of the test. The RISB consists of 40 sentence stems which examinees are asked to complete to create statements which reflect their feelings about themselves and others. The test can be administered to any size group and little experience or training is required. It has been suggested that the RISB is probably most effective in the early part of assessments, to give the examinees some distance from the examiner to facilitate their expressing themselves more openly.

Most analyses of the RISB are not objective, but subjective, relying on the clinical expertise of the evaluator. See Ackerman, supra, at 287. The manual suggests that content analysis of the RISB be similar to that done with the TAT, involving the formation of hypotheses about the individual which are to be compared with other data prior to drawing conclusions. Id.

The RISB has been successful at identifying people who are depressed, who are anxious and/or who have a poor self image. Id. at 288. The RISB has not been found to be especially effective at assessing changes in psychological adjustment over the short term via pre and post test differences. Id. Note that validity data are largely based on formal scoring of the RISB, while in clinical practice formal scoring is rarely done. Id. This raises a question about the

validity of a particular interpretation, since it is based on the skill of the clinician doing the interpretation. Id.

(2) The Rohde Sentence Completion Method

The Rohde Sentence Completion Method (SCM) consists of 65 items plus an open-ended question at the end of the test. Rohde tried to choose each stem with great care, including comparing them with stems from other tests. She also arranged the order of the stems to try to lead the individual away from every day life toward the more inaccessible areas of personality. The open-ended question is a request to "Write below anything that seems important to you." No formal scoring system exists; therefore, each response is to be evaluated in the context of psychodynamic formulations and categories of need.

(3) The Sacks Sentence Completion Test

The Sacks Sentence Completion Test (SSCT) consists of 60 items which are relatively structured compared to most sentence completion test. Sacks also suggests that an inquiry be conducted at the end to maximize an understanding of the responses. The test is organized around 15 attitudes with four stems per attitude -- e.g., attitude toward women, toward mother, toward college, etc. The degree of disturbance regarding each attitude is rated on a three point scale, from none to severely disturbed.

(4) The Forer Structured Sentence Completion Test

The Forer Structured Sentence Completion Test (Forer) consists of 100 items with instructions which emphasize speed. The stems are structured to try to force the respondent to offer material useful for diagnosis, e.g., to indicate when he or she was most depressed. There are separate male and female forms. While the sentences are not formally scored, Forer suggests organizing them into seven major areas, and orders a checklist for that purpose.

(5) The Miale-holsopple Sentence Completion Test

The Miale-Holsopple Sentence Completion Test consists of 73 items and is designed to be minimally structured. The stems consist of relatively general items, rather than obviously personal ones, with the intent of minimizing the degree to which people feel threatened or exposed. The instructions are also more general than most, asking the individual to "complete each sentence in whatever way you wish." There is no formal scoring system. Instead, the authors urge the clinician to form hypotheses sentence by sentence, and to combine them into a global description after all sentences are read. To facilitate identification of unusual responses, numerous examples of characteristic responses are given for each stem. "Overall, the Miale-Holsopple is the least structured of the sentence completion tests and depends on the subjective interpretive skills of the clinician more than any of the other tests.

(6) Incomplete Sentences Task

A more recently developed instrument is the Incomplete Sentences Task. See Anastasi, supra, at 600. In its development, this instrument combined the projective approach with standard psychometric procedures of test construction and evaluation. Available in a School Form (Grades Seven to Twelve) and a College Form, the test yields three scores: hostility, anxiety, and dependence. These constructs were chosen on the basis of personality theory and because of their clinical importance to any adjustment of student.

As a psychometric instrument, this test would be strengthened with more representative norms and with further research on reliability, validity and interrelation of the three scores. Initial results are promising, however, and provide more objective data than are customarily available for projective instruments.

c. Acceptance by Mental Health Professionals

In a report presented at the meeting of the American Psychological Association in 1983, psychologist B. Luben, R. Larsen and J. Matarazzo indicated that "Sentence completion tests (all kinds) rank 7.5 among psychological tests in their survey of psychologists." The RISB by itself ranked twelfth and is the only sentence completion test mentioned by name in their survey. Sentence completion tests also ranked 7.5 in a 1982 survey, up from 8.5 in a 1969 survey. The RISB also ranked twelfth in a 1982 survey, down from 10th in a 1969 survey but up from 61st in a 1959 survey. [The MMPI was the number one ranked test in this survey of psychologists.]

4. Projective Drawings

Most psychologists recognize that not all the information desired in psychological investigations can be gained through verbal interaction. Ackerman, supra, at 248. Children start drawing before they learn to talk, therefore drawing techniques can be very useful in assessing a child's intellectual and personal functioning. Id. The following tests have been used in the area of projective drawing and could be encountered in divorce litigation.

a. Types of Projective Drawing Tests

(1) DRAW-A-MAN TEST

Goodenough developed the Draw-a-Man test in 1926 for children between the ages of three and fifteen as a quick estimate of the child's intellectual development. Id. at 249. Goodenough developed a 50-item scoring scale. Id.

(2) THE GOODENOUGH-HARRIS DRAWING TEST

In 1963, Harris updated the Goodenough Draw-a-Man test to include a drawing of the self and of a woman. Id. Scoring criteria were presented for each of the drawings. Harris added 22 items to Goodenough's original criteria. Id. Sattler, one of the foremost experts in intelligence testing, points out that the "Goodenough drawing test is an acceptable screening instrument for use as a non-verbal measure of cognitive ability, particularly with children under 10 or 11 years of age."

Id. Sattler felt however that the drawing test should not be used as the only measure of intelligence due to reduced validity. Id.

(3) HOUSE-TREE-PERSON DRAWING TEST

In working with non-compliant children, Buck, out of desperation, requested subjects to draw a house, a tree, and a person. Id. He chose these three items, as they represented the items most frequently spontaneously drawn by children. Id. The subject is asked to draw a house, a tree, and a person on separate pages. After the drawing of the person is completed, the subject is asked to draw a person of the opposite sex. Id. Although a thorough discussion of the interpretation of the house-tree-person is not possible in a manuscript of this nature, "the house is thought to represent the environment, the tree depicts growth, and the person represents the integration of the subject's personality." Id. at 250.

(4) DRAW-A-FAMILY TEST

The Draw-a-Family test was originally developed by Appel and later elaborated on by Wolff. Id. There are generally two acceptable methods of presenting this test. Id. One method requires the subject to "draw a picture of your whole family, " and the other requests the subject to "draw a picture of a family." Id. Some examiners prefer the more ambiguous request to draw a family as it provides the examiner with the opportunity to measure the subject's identification with their own family. Id. Family constellation, placement of individuals in the family, relative sizes of the family members and other variables are used in interpreting the draw-a-family test.

(5) KINETIC FAMILY DRAWING

The Kinetic Family Drawings (KFD) test was developed by Burns & Kaufman in 1970. Id. "The approach of using kinetic (action) instructions - i.e., asking the child to produce a drawing of figures moving or doing something has been found to produce much more valid and dynamic material in the attempt to understand the psychopathology of children in a family setting." Id.

The instructions given by the examiner state, "I would like you to draw a picture of everybody in your family doing something." Id. Of particular note in the interpretation of this picture is whether the family is doing something together, or each individual is doing something separately. Id.

In the context of a custody evaluation, the KFD can be very enlightening as a way of understanding the child's perception of his family. Evaluators usually interpret both the general themes and dynamics represented, and sometimes the difference in physical distance between the child and each parent as an indication of relative emotional closeness or distance. Some recent research has begun to provide evidence of a correlation between the distance between figures on the KFD and the child's perceived closeness to the people represented in the drawing, but strong confirmation of even that interpretation of the KFD awaits the accumulation of more data.

As with any psychological test used in the context of a custody evaluation, each of these four tests has some utility. The most reasonable use of these tests' results is to serve as one additional source of information to be considered in the context of **all** the facts in making recommendations regarding custody and access. Even the Bricklin, the only test of the group specifically designed for custody evaluations, does not have enough validation data to justify its use as the **only** test given in a particular case.

(6) OTHER DRAWING TECHNIQUES

There are many minor drawing techniques that have been developed for use in interpreting specific kinds of personality concerns. Id. The number of drawing techniques developed only by the creativity of the examiner. Id. Examples include, (1) The Draw-a-Person-in-the-Rain Test to measure how the individual deals with unpleasant environmental stress; (2) The Draw-a-Dream Test was designed to depict a dream that the child actually has had or one that the child would like to have; (3) The Representational Family Drawing which was developed by Oaklander in an effort to measure how the child representationally perceives each family member. Id. at 251. The instruction of "draw a picture of something that represents each member of your family" is given. Id.

b. Cross-Examination of the Projective Test Expert

Projective tests, because of their subjective interpretation, are topics for intense cross-examination. If projective tests are the only tests administered, expose to the jury insufficient data exists to base his conclusions.

Some of the questions which might aid the attorney in cross-examination of the psychological expert, are as follows:

(1) Were projective techniques administered?
(2) If the psychologist used only projective methods for testing, why?
(3) Were the instructions for the specific projective tests followed according to the manual? Ask the examiner to repeat the instructions that were given.
(4) Were the projective instruments administered prior to objective instructions? If not, why not?
(5) What projective drawings were used?

Lynn Little St. Leger, *Psychological Examinations*, Advanced Family Law Course BB, 99-100 (1990).

c. Reliability and Validity

In general, projective instruments are less suspectable to faking than are self-report inventories. Anastasi, supra, at 613. The purpose of projective techniques is usually disguised. Id. Even if an individual has some psychological sophistication and is familiar with the general nature of a particular instrument such as the Rorschach or TAT, it is still unlikely that he or she can predict the intricate ways in which these responses will be scored and interpreted. Id. Moreover, the respondent soon becomes absorbed in a task and is less likely to resort to the customary disguises and restraints of interpersonal communication. Id.

On the other hand, it cannot be assumed that projective tests are completely immune to faking. Id. Several experiments with the Rorschach, TAT, and other projective instruments have shown that significant differences do occur when respondents are instructed to alter their responses so as to create favorable or unfavorable impressions or when they are given statements suggesting that certain types of responses are more desirable. Id.

[It is obvious that most projective techniques are inadequately standardized with respect to both administration and scoring. Id. at 614. Yet evidence exists that even subtle differences in the phrasing of verbal instructions and an examiner-examinee relationships can appreciably alter performance on these tests. Id. Even when employing identical instructions, some examiners may be more encouraging or reassuring, others more threatening, owing to their general manner and appearance. Id. Such differences may affect response productivity, defensiveness, stereotyping, imaginativeness, and other basic performance characteristics. Id. In the light of these findings, problems of administration and testing conditions assume even greater importance than in other psychological tests.] Id.

Equally serious is the lack of objectivity in scoring. Id. Even when objective scoring systems have been developed, the final steps in the evaluation and integration of the raw data usually depend on the skill and clinical experience of the examiner. Id. But perhaps the most disturbing implication is that the interpretation of scores is often as projective for the examiner as the test stimuli are for the examinee. Id. In other words, the final interpretation of projective test responses may reveal more about the theoretical orientation, favored hypothesis, and personality idiosyncrasies of the examiner than it does about the examinee's personality dynamics. Id.

Another common deficiency to many projective instruments pertains to normative data. Id. Such data may be completely lacking, grossly inadequate, or based on vaguely described populations. Id. In the absence of adequate objective norms, the clinician falls back on his or her "general clinical experience" to interpret projective test performance. Id. But such a frame of reference is subject to all the distortions of memory that reflect theoretical bias, preconceptions, and other idiosyncrasies of the clinician. Id. The clinician may, thus, lack sufficient first-hand familiarity with the characteristic reactions of normal people. Id.

In view of the relatively under-standardized scoring procedures and the inadequacies of normative data, score reliability becomes an important consideration in projective testing. Id. at 615. Interpretative score reliability is concerned with the extent to which different examiners attribute the same personality characteristics to the respondent on the basis of their interpretations of the identical record. Id. Few adequate studies of this type of score reliability have been conducted with projective tests. Id. Some investigations have revealed marked divergences in the interpretations given by reasonably well qualified test users. A fundamental ambiguity in such results stems from the unknown contribution of the interpreter's skill. Id. Neither high nor low score reliability can be directly generalized to other scores differing appreciably from those utilized in the particular investigation. Id.

Retest reliability also presents special problems. Id. at 616. With long intervals, genuine personality changes may occur which the test should detect. Id. With short intervals, a retest may show no more than recall of original responses. Id. A large chance of variations are to be

expected under such circumstances. Id. Ratios and percentages computed with such unreliable measures are often even more unstable than the individual measures themselves. Id. For any test, the most fundamental question is that of validity. Id. Many validation studies of projective tests have been concerned with concurrent criterion related validation. Id. Most of those have compared the performance of contrasted groups, such as occupational or diagnostic groups. Id. The large majority of published validation studies on projective techniques are inconclusive because of procedural deficiencies in either experimental controls or statistical analyses, or both. Id. at 617. (This is especially true of studies concerned with the Rorschach test. Id.) Some methodological deficiencies may have the effect of producing spurious evidence of validity where none actually exists. Id. The customary control for the types of contamination in validation studies is to utilize blind analysis, in which the test worker is interpreted by a scorer who has had no contact with the respondent and who has no information about her or him other than that contained in the test protocol. Commissions have argued, however, that blind analysis is an unnatural way to interpret projective test responses and does not correspond to the ways these instruments are used in clinical practice. Id.

Besides their questionable theoretical rationale, many projective techniques are found clearly wanting when evaluated in accordance with test standards. Id. at 621. This apparent contradiction can perhaps be understood if we recognize that, with a few exceptions, projective techniques are not truly tests. Rather than being regarded and evaluated as psychometric instruments, or tests in the strict sense of the term, projective techniques are coming more and more to be regarded as clinical tools. Id. Thus, they may serve as supplementary qualitative interviewing aids in the hands of a skilled clinician. Their value as clinical tools is proportional to the skill of the clinician and, hence, cannot be assessed independently of the individual clinician using them. Id. at 622. Attempts to evaluate them in terms of the usual psychometric procedures would be inappropriate. But by the same token, the use of elaborate scoring systems that yield quantitative scores is not only wasteful but also misleading. Id. Such scoring procedures lend the scores an illusory semblance of objectivity and may create the unwarranted impression that the given technique can be treated as a test. Id. The special value that projective techniques may have is more likely to emerge when they are interpreted by qualitative clinical procedures than when they are quantitatively scored and interpreted as psychometric instruments. Id.

V. New Developments in Psychological Testing Within the Last Decade

A. Bricklin Perceptual Scales (Bps)

1. History, Background and Purpose of Bricklin

The Bricklin Perceptual Scales, developed by Barry Bricklin, Ph.D., were first published in 1984. Although only recently published, the research, testing and validation of the test has been an ongoing process since the early 1960's. The test was designed to measure "a child's unconscious or nonverbal perception of each parent in the areas of competence, supportiveness, follow-up consistency, and possession of admirable traits." (B. Bricklin, Bricklin Perceptual Scales Manual, at 6.)

Dr. Bricklin developed the test because, although he is an "expert" in the use of the standard projective tests, he believed that these tests were not well-suited for use in custody disputes. Bricklin concluded that the following characteristics were required of any test which was to be useful to a court in custody disputes:

(1) The test would have to be based on the child's perceptions, rather than on parental behavior.
(2) The test would have to look at the child's observations, thoughts and feelings about each parent.
(3) The test would have to rate each parent in each area covered by the test.
(4) The test would have to avoid asking the child direct questions.
(5) The test would have to avoid asking the child to make a direct choice between parents.
(6) The test would have to be capable of reflecting a particular need of a child that is so compelling that the parent who could best satisfy that need would receive recognition of it in the scoring scheme.
(7) The test would have to be based primarily on unconscious, rather than conscious, responses. (Id. at 12 - 13.)

For valid and reliable use, Bricklin suggests that the test be used only on children ages six years and up.

2. Construction of BPS

Four areas are tested on the BPS: (1) the child's perception of a parent's competence, (2) supportiveness, (3) follow-up consistency, and (4) possession of admirable traits (e.g., keeping promises, trustworthiness, altruism, ability to accept criticism, etc.). (Id. at 17.)

The test consists of 64 questions, 32 of which pertain to the child's perceptions of m other and 32 to the child's perceptions of father. First, the child is asked to respond to the questions verbally, and the answer is recorded. The child is then asked to respond to the same questions non-verbally on a continuum line which ranges from Very Well to Not So Well. (Id. at 17.)

Each question is presented by showing the child a card, which pictures a line with the words "Very Well" at the extreme right and "Not So Well" to the left. (A few cards say "Very Often" and "Not So Often"). (Id. at 18.)

After the child has been asked the questions verbally and they have been recorded, the nonverbal portion of the test begins. The card is held in front of the child and clinician asks, "If this (pointing to the end of the line marked "Very Well") is Dad doing very well and helping you calm down, and this (pointing to the end marked "Not So Well") is Dad doing not so well at helping you calm down, where on this line would Dad be?" The child then uses a stylus to punch his or her response along the continuum. (Id. at 18.) The questions, as well as the scoring grid, are printed on the reverse side of the card. The grid ranges from 1 to 60, with 1 being "not so well" and 60 being "very well."

It is assumed that the child's placement on the continuum represents an unconscious, gut-level response which reflects a combination of observations, thoughts and feelings "weighted" by the child's own value system of the parent being considered in the area covered by the card. (Id. at 19.)

The following are examples of the questions presented to the child in the BPS:

(a) COMPETENCY

Verbal:

"If you had questions bout where babies come from (children under 7 and older) sex and babies (children 8 and older) how helpful do you think you would find it to talk to Mom about this?"

Non-Verbal:

"If this is Mom being very helpful at answering your questions, and this is Mom being not so helpful at answering your questions, where on this line would Mom be?"

(b) SUPPORTIVENESS

Verbal:

"If Dad happens to be the one to take you to the doctor, how is he at making you feel comfortable about this?"

Non-Verbal:

"If this is Dad doing very well at helping you to feel comfortable, and this is Dad doing not so well at helping you to feel comfortable, where on this line would Dad be?"

(c) FOLLOW UP CONSISTENCY

Verbal:

"How often does Mom make sure you do your jobs around the house?"

Non-Verbal:

"If this is Mom very often making sure you do your jobs around the house, and this is Mom not so often making sure you do your jobs around the house, where on this line would Mom be?"

(d) ADMIRABLE TRAITS

Verbal:

"When Dad has a chance to spend some time with other people, how much does he seem to enjoy this?"

Non-Verbal:

"If this is Dad very much enjoying time with other people, and this is Dad not so much enjoying time with other people, where on this line would Dad be?" Bricklin Perceptual Scales, Cards 10, 17, 41 and 56.

3. Supplements to the BPS

To date, Dr. Bricklin has updated the BPS with six short supplements, which are delivered along with the test. A brief synopsis of the supplements follow:

(a) Suspect Reactions

Where the child may prefer one parent as caretaker over the other, but the evaluator feels that the preference may not be in the child's best interest, Bricklin admonishes the test giver to be wary of suspect reactions. Bricklin outlines the following as suspect regarding a child's conscious or verbal choice:

a. it is based on what the child has been told rather than actual experience.
b. the evaluator suspects that the child is trying to "save" a parent, for fear of that parent falling apart if not chosen.
c. the evaluator senses the possibility of the child having been manipulated, through gifts, promises, or special treatment (i.e., lack of rules, letting the child stay up past bedtime, etc.). BPS Manual Supplement No. 1 6/11/85.

(b) Close Calls

Supplement 2 helps prepare the evaluator for cross-examination on close calls, where the contestants have come out fairly equally on test results. To try to further assess close calls, Dr. Bricklin suggests that the evaluator follow up with cards 4, 6BM, 7BM, 7GF, and 10 of the Thematic Apperception Test (TAT). BPS Manual Supplement No. 2 9/12/85.

(c) Reports

Bricklin bases his own reports on a combination of the BPS, the Perception of Relationships Test (commonly referred to as the PORT), it involved drawing tests), My Parent Would, and the House-Tree-Person Test. He also sometimes uses the Rorschach. When faced with a real adversary case, where one side wants to litigate regardless of the data, Bricklin advocates writing a short report, tying down the pertinent date to prevent "fishing expeditions." BPS Manual Supplement No. 3 3/28/86.

(d) Additional Directions for Child With Conscious Bias

In the most recent supplement, Bricklin addresses the "MMU." If a child has a conscious bias, or is a child with "mind-made-up" (MMU), that child has probably been programmed or bribed. Bricklin suggests the following as tools to ensure accurate BPS data with MMU children:

a. Speak very slowly when you are reading the questions to ensure that the child understands and occasionally ask the child to repeat what was said.
b. Challenge some responses, especially those given in haste. Ask for examples. If the child cannot give any example to back up the quick response, the response may have been contrived.

c. Repeat critical words in long questions.
d. Watch involuntary hand movements - does the child have an initial impulse in one direction on the continuum and then force himself to respond on the other end of the continuum?
e. Does the child avoid eye contact? Resist humor or friendliness? BPS Manual Supplement No. 6 6/30/88.

4. Scoring the BPS

When 64 cards are completed, there will be 32 paired scores, 11 of which measure Competency, 11 measure Supportiveness, 3 measure Follow-Up Consistency and 7 measure Admirable Traits.

The clinician goes through the paired Point Scores for each item and draws a circle around the higher number. The clinician then counts the number of items on which Dad had the highest score and the number of items on which Mom had the highest score. The parent achieving the higher number of highest scores perceives that parent as the primary caretaking parent. (Id. at 33 -4.)

5. Reliability and Validity of the BPS

Bricklin, in dealing with reliability, notes that "there are no reasons to expect the measurements reported here to exhibit any particular degree of stability, since they should vary in accordance with changes in the child's perceptions." Ackerman, supra, at 282, citing Bricklin Perceptual Scales, at 36. Limited reliability studies have been conducted, one in particular, where 12 children in custody cases were retested on the BPS within a 7-month span of the original testing. None of the data changed significantly. In another study, six children in a non-adversary population were retested, and the only change was for a child who was in family therapy, an expressed purpose of which to increase the quality of parenting of the new custodial parent. Id.

Validity has been tested in several ways. Id. One method involved validation against a previously validated instrument, the Perception of Relationships Test (PORT). Note: the validity method of both Bricklin Perceptual Scales and the ASPECT test should be questioned as one method of validity which appears to be highly touted by the authors of these tests is the fact that Judges picked the same person, a primary caretaker of the child in a custody suit, as the person chosen by ASPECT or Bricklin, or the PORT test. Knowing the adversarial system as well as we all do, much depends upon the quality of lawyering and the way the case was handled in situations where custody is determined by the Court. The author of this paper holds certain reservations about this particular method of validity. However, if the validity data included several hundred cases, those concerns could be eliminated.

The second method of validity measure utilized in Bricklin's Perceptual Scales involved two kinds of child questionnaires in which each child was asked to name the parent more likely to lend assistance, support, or control in a wide variety of circumstances, and one which required

the child to describe what each parent would do in prescribed practical and fantasy situations. These last two methods had a 70% and 87% validity, respectively. Bricklin Perceptual Scales, at 43.

Bricklin acknowledges that some experts have not been gathering verbal responses due to their reasoning that forcing the child to give a verbal response may cause tension if the nonverbal response is different. Bricklin asserts that BPS responses are rarely "fake" because "BPS nonverbal responses allow children to tell what they want to tell without ever having to come into conflict with parental verbal commands." BPS Manual Supplement No. 4 10/15/86.

Bricklin admonishes evaluators who encounter a child who robotically punches one end of the continuum or the other to instruct the child as follows: "We already know you want to live with (whoever). This test is designed to give us other information. Spread your responses out and give honest answers." BPS Manual Supplement No. 4 10/15/86

Some interesting validity research has been done in the short time that the Bricklin has been out. Twenty-nine out of thirty-six cases in which the test as given went to Court to determine custody. In twenty-seven of the twenty-nine cases, the judges, using all evidence presented, including life histories, vocational data, school and medical records, testimony of friends and relatives, etc., gave custody to the same parent as the BPS scored as being the primary caretaker. (Id. at 33-4.)

In the supplement, Bricklin notes that in formal hearings, the judge's choice and the BPS results of who would be the better primary caretaking parent had an agreement rate of 89 percent. The same agreement rate was arrived at by comparison of results on other psychological tests and the BPS. The agreement rate between the psychologists's interpretations and the clinical life history data was 91 percent. The agreement rate between the primary caretaking parent and all available tests was 97 percent. The survey was based on 27 respondents, using a sample pool of 141 cases. BPS Manual Supplement No. 5 2/15/88.

6. Advantages of the BPS

The Bricklin Perceptual Scales are so new that it is not easy to assess its advantages since the follow-up research data on this test is so limited, unlike, for example, the MMPI or the Rorschach. However, it is clear that the Bricklin is much more of a valid tool for the clinician who is faced with evaluating one or both parents and making a recommendation to the Court. Other personality tests may reveal that Dad is an alcoholic, for example, but it does not reveal how this trait in Dad impacts on his child, or if it is even relevant. Additionally, what has an impact on one child does not necessarily have impact on another.

Dr. David Brodzinsky, Associate Professor of Clinical and Developmental Psychology at Rutgers University, states the following about the BPS:

> Despite the limitations of these alternative assessment procedures [such as the BPS], they represent a valuable addition to the field of child custody evaluation. Most important, they shift the focus from a more traditional clinical assessment to one in which the evaluator is focusing more on a functional analysis of the parties' competencies within specific child care roles. As such, these instruments are likely to provide information that

is particularly relevant to the issues before the court. Future research, however, will need to resolve the dilemma regarding validation of these measures.

David M. Brodinsky, "On the Use and Misuse of Psychological Testing in Child Custody Evaluations," PROFESSIONAL PSYCHOLOGY: RESEARCH AND PRACTICE, vol. 24, no. 2, 1993. p. 214 - 219.

As the BPS and other tests become more frequently used, their reliability and validity in custody evaluations will become more clear. To date, the Bricklin Perceptual Scales and the Perception of Relationships Test (PORT) have been administered more than 50,000 times as part of a custody evaluation. They have been used in all fifty states and accepted as part of courtroom proceedings in these states. They have been purchased by every level of government from county to the United States government. Moreover, hundreds of major hospitals and universities use the BPS and PORT in their outpatient clinics.

Dr. Bricklin believes that the BPS successfully addresses all the main complexities which plague custody decision-making. (Id. at 16.) At the very least, it is a start at addressing what is really relevant in custody disputes: **the child**.

7. Cross-Examination of the Expert

Any cross-examination of the expert should elicit the information that the BPS is a very new testing device and has not yet received outside confirmation of reliability and validity.

The attorney should pay attention to the age of the child that the BPS is given to, because even Dr. Bricklin suggests that the "bottom line" age, to insure for reliable and valid use, to be age six and above. If the child is eight years of age or below, the attorney should ask the psychologist if developmentally the child is capable of abstract thinking. Additionally, the test may not be very reliable for children over the age of 12, because if that child is campaigning to be with a particular parent, the questions are structured in such a way that a bright 12 year-old would realize the result of his or her answers and, therefore, answer in such a way to make that particular "win." Ask the psychologist if it is possible that this has happened.

Additionally, if video taping the administration of the test is not possible, then a detailed examination of whether the clinician could, in any way, have influenced the child's answers should be undertaken. Depending on how the questions are asked the clinician, in administering the test, may greatly influence the child's answers.

B. Sex Abuse Legitimacy Scales (Sals)

1. Parental Alienation Syndrome

The Sex Abuse Legitimacy Scales were developed by Richard A. Gardner, M.D. Dr. Gardner, a practicing child psychiatrist and adult psychoanalyst, is a professor at Columbia University in New York. Dr. Gardner has published extensively on the subject of psychotherapy

with children of divorce. In his book, The Parental Alienation Syndrome and the Differentiation between Fabricated and Genuine Sex Abuse, Dr. Gardner discusses the development and underlying theory supporting his Sex Abuse Legitimacy Scales (hereinafter referred to as "SALS"). [R. Gardner, The Parental Alienation Syndrome and the Differentiation Between Fabricated and Genuine Sex Abuse, (New Jersey: Creative Therapeutics, 1987).]

Dr. Gardner developed the SALS to focus on and battle what he has dubbed the "parental alienation syndrome." (Id. at 67.) Dr. Gardner defines "parental alienation syndrome" as:

> . . . a disturbance in which children are preoccupied with deprecation and criticism of a parent - denigration that is unjustified and/or exaggerated. The notion that such children are merely 'brainwashed' is narrow. The term brainwashing implies that one parent is systematically and consciously programming the child to denigrate the other. The concept of the parental alienation syndrome includes the brainwashing component, but is much more inclusive. It includes not only conscious but subconscious and unconscious factors within the programming parent that contribute to the child's alienation from the other. Furthermore, (and this is extremely important), it includes factors that arise within the child - independent of the parental contributions - that play a role in the development of the syndrome. In addition, situational factors may contribute, i.e. factors that exist in the family and the environment that may play a role in bringing about the disorder." (Id. at 67-8.)

In laymen's terms, the parental alienation syndrome is when a child, for no valid reason, is distant and/or has dislike for one of his or her parents.

Dr. Gardner contends that there are two reasons for the increase in manifestations of the parental alienation syndrome: (1) the "best interest of the child" presumption superseding the sexist and outdated "tender years" presumption (i.e. the latter representing the theory that mothers should be favored over fathers as custodians of children of tender years), and (2) with joint managing conservatorships taking an upswing, parents have been more apt to "brainwash" a child to insure victory, without considering that the child may in fact be better off with the denigrated parent. (Id. at 68-9.)

Dr. Gardner notes that parental alienation will manifest itself with the "brainwashed" child developing an "obsession" about the "hated" parent. Given the proper forum i.e., the court-appointed psychologist's office, judge's chambers, or in open court, the child will provide "a command performance," giving a dissertation on the alleged wrongs committed by the "hated" parent in a speech that has a rehearsed quality. The child's alienation may be rationalized by trivial events such as the "hated" parent forcing the child to write out an incorrect spelling word 25 times to insure that the same error will not be repeated, or speaking loudly when telling the child to brush his or her teeth before bed, or telling the child to get something for the hated parent. Gardner states that:

> "The professions of hatred are most intense when the children and the loved parent are in the presence of the alienated one. However, when the child is alone with the allegedly hated parent, he or she may exhibit anything from hatred, to neutrality, to expressions of love and affection. When these children are with the hated parent, they may let their

guard down and start to enjoy themselves. Then, almost as if they have realized that they are doing something 'wrong,' they will suddenly stiffen up and resume their expressions of withdrawal and animosity. Another maneuver commonly seen in this situation is the child's professing affection to one parent and asking the parent to swear that he or she will not reveal to the other parent the professions of love. And the same statement is made to the other parent. In this way these children 'cover their tracks' and thereby avoid the disclosure of their schemes. Such children may find family interviews with therapists extremely anxiety provoking because of the fear that their manipulations and maneuvers will be divulged." (Id. at 71-2.)

Dr. Gardner notes that the proximity of the loved parent has an impact on the child's reactions, and the closer the loved parent is (in the room, outside in the waiting area, at home), the more likely the child will manifest his or her alienation. Dr. Gardner urges that anyone involved in custody litigation recognize the phenomenon. (Id. at 72.)

2. **Factors Contributing to Development of Parental Alienation.**

Dr. Gardner identifies the following four factors as contributing to the child's development of parental alienation:

(a) Brainwashing

Brainwashing includes conscious acts of programming the child against the other parent. For example:

(i) Dad convincing himself that Mom is having an affair, and telling the child about innocent conversations Mom's had with men as "proof" of Mom's infidelity.

(ii) Mom complaining to the child or children so much about new financial restrictions that the children think they may go without food, clothing or shelter because Dad's now a Scrooge.

(iii) Talking to the children about problems with the other spouse and exaggerating them, such as: if Mom has a glass of wine over a business lunch, she becomes "an alcoholic." (Id. at 75-88.)

(b) Subtle Programming

Subtle programming is akin to brainwashing, but often unconscious. An example of subtle programming would be the old guilt trip trick. Example: saying to the child, "your father has got other people in his life, all I have in the whole world is you. When you're gone, I am so alone". (Id. at 80-8.)

(c) Factors Arising Within The Child

The parental alienation often develops from contributions arising from "psychopathological factors within the child." For example: the child believing for some reason

that he or she is "bad for wanting to visit" the noncustodial parent or a daughter believing that her father must choose between her and his new girlfriend. (Id. at 89-92.)

(d) Situational Factors

Situational factors would be those ". . .external events that contribute" and/or ". . .abet the internal psychological processes in the parents and in the child." (Id. at 92.) Examples of situational factors that add to the parental alienation syndrome would be when one sibling observes another being punished, treated badly, or rejected for speaking out on behalf of the rejected parent. (Id. at 92-6.)

3. Fabricated and Bona Fide Sex-abuse Allegations in Custody Disputes

Dr. Gardner discusses the changes of the late '70's and early 80's, of courts shifting away from the tender years presumption toward joint custody, whereby "an attempt was made to provide a more egalitarian role for both parents in their children's upbringing." (Id. at 99.) Dr. Gardner continues that perhaps because of "a progressive erosion of the mother's secure position in custody disputes . . .[i]n 1982 or 1983 I began seeing a new development, namely, the utilization of fabricated allegations of child sex abuse in the context of custody disputes." (Id. at 100.)

Dr. Gardner acknowledges that children have been exposed to sex abuse as an increasingly common topic on television, and that sex abuse prevention programs have become standard in more schools, beginning even at the nursery school level. Both parents and children are educated and have the idea that an allegation of sex abuse will certainly get the Court's attention in a custody suit. As Dr. Daniel C. Schuman discussed at an annual conference of the American Academy of Psychiatry and Law in 1984, "[h]eightened instinctual forces in children and regressive loosening of pre-litigation character defenses in adults, both in the context of stressful family breakdown, combine to generate genuine perceptions of abuse but invalid reports." (D. Shuman, "False Accusations of Physical and Sexual Abuse," Annual Conference of American Academy of Psychiatry and the Law, Nassau, The Bahamas, October 26, 1984.)

a. Criteria for Construction of the SALS

Dr. Gardner propounds the following criteria for assessing the child, the father, and the mother:

1. THE CHILD

Dr. Gardner identifies ten criteria as being very valuable in differentiating between genuine and fabricated sex abuse when interviewing the child.

(i) Presence of Parental Alienation Syndrome.

Fabricating children are thought to be more likely to manifest symptoms of parental alienation syndrome.

(ii) Receptivity to Divulgence .

Kids who are genuine tend not to be thrilled by the prospect of spilling their guts to case workers, psychologists, lawyers, judges, etc. The opposite is true with fabricators. Fabricators will talk, and are often encouraged to do so by the accuser, to anyone who will listen.

(iii) Providing Specific Details .

Gardner stipulates that children who have genuinely been abused will present more concrete facts than children who are fabricating.

(iv) Credibility of the Description.

If a child is making things up, it shows in his or her description.". . . [I]t is in the description of the ejaculate, especially, that the fabricator is likely to provide preposterous explanations." (Gardner, supra, at 111.) Moreover, the fabricating child is likely to take the interviewer's lead, and describe the "stuff" that came out of perpetrator's penis as yellow and clear like urine -- the only other "stuff" with which the kid has a frame of reference.

(v) Guilt Relating to the Consequences of the Disclosure of the Accused.

Fabricators do not tend to exhibit guilt or remorse over what might happen to the perpetrator as a result of the disclosure. Children who have truly been abused "may feel guilty over their disloyalty and the recognition that the disclosure is going to result in formidable painful consequences for the perpetrator.

(vi) Guilt Relating to Participation in Sexual Activities.

The fabricator does not feel guilt over the sexual activities he or she has allegedly taken part in. The child who has truly been abused may feel guilty as a result of feeling pleasure over said activities.

(vii) Fear of the Alleged Perpetrator .

The fabricator usually is not frightened of the alleged perpetrator -- only of what the accused might do to the child as a result of the false accusation.

(viii) Sexual Excitation .

Children who have truly been abused experience an early "turn on" for which they may seek an outlet, such as physical contact with the interviewer.

(ix) Desensitizing Play .

Children who have genuinely been abused may engage in desensitizing play -- such as ejecting the doll that represents the perpetrator from the doll house.

(x) Attitude Toward Genitalia .

Victims of abuse tend to consider their genitals as having been damaged. Fabricators do not have such feelings, and will not independently characterize their genitals as damaged. (Id. at 109-17.)

Of moderate significance to Gardner are the following:

(i) the litany of fabricator has a rehearsed quality;

(ii) genuine victims frequently manifest depression;

(iii) genuine victims tend to withdraw from involvement with others;

(iv) sexually abused children are often compliant, developing a cheerful facade to ward off the threat of the dire consequences outlined by the perpetrator for noncompliance;

(v) fabricators borrow their scenarios from other experiences, such as classroom presentations, movies, etcetera;

(vi) victims tend to be more likely to suffer from psychosomatic disorders;

(vii) children who have been abused tend to display regressive behavior; and

(viii) children who have been abused often display a deep-seated sense of betrayal. (Id. at 117-21.)

Dr. Gardner recognizes that sleep disturbances, chronicity, pseudo-maturity, seductive behavior with the perpetrator and retraction may all potentially have high value in differentiating genuine from fabricated sex abuse. (Id. at 121-24.)

2. THE MOTHER

(i) Initial Scenario

The interviewer is to determine how mama first learned of the abuse -- did the child report that dad attacked him in the shower, or did Mom, for example, inquire whether Dad had washed his penis and if his penis "got a little hard when Daddy was washing it?"

(ii) Shame

Gardner notes that many mothers would be ashamed at having to report reprehensible conduct on the part of their husbands, if in fact the husband has truly abused the child.

(iii) Seeking Hired Gun Evaluator

Gardner thinks that mothers of fabricators will, in the words of one of Dallas' top family litigators, "hire her a whore." Rather than trusting the court-appointed psychologist, the mother of the fabricator will bring in a hired gun.

(iv) Joint Interview Corroboration.

Gardner has noted that fabricators, during the course of joint interviews with Mom, will frequently give Mom side-long glances to "check" their stories.

Gardner points out the following criteria as having a moderate affect on differentiating mothers of children fabricating sex abuse as opposed to mothers of children who have been genuinely abused:

(i) The mother of a genuine victim will appreciate the psychological trauma suffered by the child as a result of repeated interrogations;

(ii) Mothers of children who have been abused may still recognize the importance of the father-child relationship, and seek to preserve it in spite of the abuse;

(iii) Gardner argues that mothers of fabricators were not typically abused as children, but will take into account the childhood history of sex abuse of the parties;

(iv) While mothers of children genuinely abused tend to be passive or incapacitated, according to Gardner, mothers of fabricators tend to be aggressive and outspoken. (Id. at 124-31.)

Gardner observes that the personality characteristics regarding the mother of the child alleging abuse should also be assessed. (Id. at 131-32.)

3. THE FATHER

While Gardner admits that generalizations regarding the personality characteristics of fathers who are bona fide abusers are dangerous, he outlines the following as very valuable differentiating criteria to distinguish a father of a fabricator from a father who has abused his child:

(i) Bribes or Threats.

Children who have truly been abused have frequently been bribed or threatened by the erring father.

(ii) Indignation.

Fathers who have been falsely accused suffer from extreme indignation.

(iii) Presence Of Other Sexual Deviations.

Gardner argues that a man who engages in pedophilia will exhibit other deviations such as "exhibitionism, rape, voyeurism, sado-masochism, and homosexuality." Here

Gardner displays one of his biases: he considers what he characterizes as "obligatory homosexuality" as deviant. (Id. at 132-34.)

Gardner points out the following criteria as having a moderate affect on differentiating fathers of children who have been genuinely abused:

(i) Fathers who have a history of having been abused themselves are more likely to abuse their children;

(ii) Fathers of fabricators are more likely to be enthusiastic about taking a lie detector than someone who has abused their child;

(iii) Fathers who abuse are more likely to have a history of drug and/or alcohol abuse;

(iv) Fathers who abuse are more likely to suffer from low self-esteem than fathers who do not;

(v) A father who has abused his child will be more likely to regress in stressful situations;

(vi) An abuser may choose a career that brings him into contact with children; and

(vii) Fathers who abuse their children are more likely to be social isolates. (Id. at 134-37.)

Gardner observes that stepfathers are more likely to abuse than natural fathers. He further notes that fathers who abuse, often rigid and strict, tend to be very moralistic about sex. (Id. at 137-38.)

b. Clinical Evaluation

Dr. Gardner advocates the use of a clinical evaluation in conjunction with the SALS. Dr. Gardner employs the use of audiocassettes and videocassettes to document the evaluation, and never agrees to not being allowed to interview the child in the presence of the alleged perpetrator. In fact, he mandates a court order allowing him to interview all of the parties, in any sequence he feels is warranted. Prior to interviewing a particularly young child, he suggests that the interviewer inquire of the parents pet names that may be used for body parts. Gardner suggests a direct verbal inquiry with a child, giving the child an opportunity to draw a picture and tell a story about it, and asking the child to draw a picture of a person, another of a person of the opposite sex, a third of a family, and finally, he requires the child to tell a story about the family picture. Unless the evidence or clinical examination warrants it, Gardner states that the

use of anatomically correct dolls is simply "loading the dice," making information elicited through the use of such dolls less credible. (Id. at Addendum III)

Dr. Gardner also suggests that the court-appointed psychologist research the medical and/or hospital file on the child, as by the time the evaluator is called in, physical manifestations may no longer be present.

c. The SALS Test

Many psychologists feel that Gardner's Sex Abuse Legitimacy Scale is not a "test." The reason: lack of empirical data. Dr. Gardner comes to his conclusions based solely on his clinical experience as opposed to any studies. Therefore, the SALS is best characterized as a clinical tool. Moreover, the SALS includes the following warning on the front of the instrument:

> **"WARNING**: In order to be used in a meaningful way, this instrument **must** be used in association with the information provided by Dr. Richard A. Gardner in chapters 3, 4, and 5 of his book, The Parental Alienation Syndrome and the Differentiation Between Fabricated and Genuine Child Sec Abuse (Cresskill, New Jersey: Creative Therapeutics, 1987). The book explains how best to evaluate and score each of the items in the scale. Failure to use these guidelines may result in misleading or erroneous conclusions." (Id., at Addendum III, 3.)

In the instructions to the SALS, Gardner outlines the following salient points:

". . .When medical evidence is not present, the SALS Scores may be the primary sources of information about whether the sex abuse allegation is valid." (Id.)

1. ITEMS

The SALS consists of 26 items to be propounded to the child who alleges the sex abuse; 11 items to be propounded to the accuser, especially where the accuser is the mother; and 13 items is to be propounded to the accused. The items and criteria are outlined above, but as an example, one very important item regarding the child alleging the sex abuse would be the item "very hesitant to divulge the sexual abuse," to which the interviewer would check "yes, " "no," or "not clear or not applicable."

2. CRITERIA

As outlined above, Dr. Gardner has created three categories:

(a) Very Valuable Differentiating Criteria
(b) Moderately Valuable Differentiating Criteria
(c) Differentiating Criteria of Low But Potentially Higher Value

3. SCORING

Dr. Gardner explains the use and scoring of the items as follows:

> "The items are worded so that the greater the number of **Yes** answers, the greater the likelihood that the sex abuse is genuine. In contrast, the smaller the number of **Yes**

answers, the greater the likelihood the sex abuse has been fabricated. The differentiating criteria are divided into three categories, from the most to the lease valuable. In order to give greater weight to the more valuable criteria, the following point scores are to be given for **Yes** answers in each of the three categories:

Part A. Very Valuable Differentiating Criteria - 3 points for each Yes answer.

Part B. Moderately valuable Differentiating Criteria - 2 points for each Yes answer.

Part C. Differentiating Criteria of Low But Potentially Higher Value - 1 point for each Yes answer." (Id.)

Separate scores are calculated for the child, who can score a maximum of 60 points, for the accuser, who can score a maximum of 27 points, and for the accused, who can also score a maximum of 27 points. A cumulative score is not computed.

In computing the scores, scores are weighted as follows for the child, the accuser and the accused, respectively, with "n" representing the number of "yes" checks, as follows:

Child

Part A: n x 3 = maximum of 39
Part B: n x 2 = maximum of 16
Part C: n x 1 = maximum of 5
Maximum total of A+B+C = 60

6 or below indicates fabrication; 7 to 29 are inconclusive; 30 suggests real abuse

Accuser

Part A: n x 3 = maximum of 18
Part B: n x 2 = maximum of 8
Part C: n x 1 = maximum of 1
Maximum total of A+B+C = 27

3 or below indicates fabrication; 4 to 13 are inconclusive; 14 suggests real abuse

Accused

Part A: n x 3 = maximum of 12
Part B: n x 2 = maximum of 12
Part C: n x 1 = maximum of 3
Maximum total of A+B+C = 27

3 or below indicates fabrication; 4 to 13 are inconclusive; 14 suggests real abuse

Dr. Gardner contends, based on his years of experience and research, that a score of 50% or more out of the maximum or more suggests bona fide sex abuse, whereas a score of 10% of the maximum or below suggests fabricated sex abuse. If one of the inconclusive score leans toward a significant score, that increases the likelihood of fabrication or abuse, depending on which end of the spectrum the score leans toward. (Id.)

THE SALS SHOULD NOT BE USED AS A QUESTIONNAIRE. "Rather, the scale should be used **after** the interviews with the child, accuser and accused have been completed. Both individual and joint interviews **must** be conducted in order to properly assess conflicting data that is often presented." (Id.)

d. Reliability/Validity of the SALS

If the SALS is a reliable test, then any two psychologists who administer it should get the same results. There are no studies that have been conducted regarding the reliability of the SALS, however, considering the subjectivity of the scoring of the SALS, it would not be surprising to find different results depending on the prejudices and bias of the individual psychologist.

As a rating system, or a clinical tool, the Sex Abuse Legitimacy Scales succeed as a valid instrument. In fact, it certainly comes closer to measuring the existence of sex abuse than any other test before it. If for no other reason Dr. Gardner should be applauded for his efforts. However, as a means of compiling data capable of statistical interpretation, the SALS fails. Until Dr. Gardner chooses to publish the empirical data he has gathered, and on which he bases the items and scoring of the SALS, there is no way to determine if the SALS truly measures whether sex abuse has occurred or not. In fact, Dr. Gardner himself now testifies that it is not a test and has withdrawn it as such.

e. Direct Examination Of The Expert

(1) Establish that the SALS is a systematic effort to quantify a very emotional situation in an objective manner.

(2) The SALS takes into account the suspicion that sex abuse may have been raised merely for litigation rather than for legitimate purposes.

(3) Establish that the SALS helps to determine if the parties and/or the child require counseling, and/or if visitation should be supervised or limited.

(4) Establish that the SALS provides a means for, and advocates, interviewing all of the parties involved in the case/sex abuse allegations.

(5) Establish that the SALS provides some means of help for the psychologist in determining whether a child has been sexually abused.

(6) Establish that the SALS strives to be fair in including the alleged perpetrator, the victim and the accuser.

(7) Establish that the SALS includes a multiplicity of factors under each area, allowing the interviewer to check "unclear or uncertain" (which is often the way things are).

(8) Establish that the SALS makes an effort at quantification by attributing numerical values to the assessments. Before the SALS, no other psychological test has provided such assessment.

f. Cross-Examination Of The Expert

(1) The SALS is not a true test. It is simply a rating scale. If the expert testifies that the SALS is a test, and not simply a clinical tool, point out the lack of empirical data validating the items or the criteria by which they are adjudged.

(2) There are no standards defined for making the scale other than those outlined in chapters 3-5 of Gardner's book. If the psychologist does not have the same amount of experience as Gardner, or if he or she did not adhere to the guidelines outlined in the book, the resulting scores should be disregarded.

(3) If one of the parties declines to be interviewed then ask the psychologist how he or she can quantify interviews if one of the parties declined to participate?

(4) Question the psychologist about Gardner's failure to show the building blocks through which he established his items, or the justification of his "1,2,3" rating scale, other than his own experience.

(5) The SALS is subjective, and one examiner's assessment could vary vastly from another examiner's assessment. Additionally, Gardner does not provide definitions for such words as "moralistic". What is moralistic to one psychologist may be blasphemy for another. Without such definitions and standards to go by, the test can most definitely be effected by the attitudes and beliefs of the examiner. The attorney should get the expert to admit to this problem with the SALS.

(6) Get the expert to admit that the rating scale forces dichotomous categories of responses where most responses should fall on a continuum (ex: does the child love his or her parents?)

(7) Additionally, get the expert to admit that Gardner's items and scores are based on opinion. Opinions are not equal to standards.

(8) Ask the expert if Gardner does not state things as universal truths that may in fact not be universally accepted truths: ex: although Gardner disclaims having sexist views, he characterizes the father as the "hated parent," and the mother as the "loved parent," based on a "reflection of" his "own observation." (Also note his opinion that "obligatory homosexuality" is

an illness. Further, Gardner characterizes the abuser by whether or not he has pedophilic characteristics, but never explains how he defines pedophilia.)

(9) The expert should be forced to admit that the characterization of certain items as very valuable, moderately valuable, or of low but capable of high importance is never justified by empirical data.

4. Phallic Plethysmograph

The Phallic Plethysmograph, affectionately also known as the "**dipstick**," "**peter meter**," and "**hardo'meter**," may well be the functional, modern-day equivalent of the Salem Water Test (used to determine if a person was a witch or warlock as alleged, if the accused drowned when submerged in the "Salem Water Test" they were obviously innocent; if they floated, they were burned at the stake as only those guilty as charged could survive the water test). The Phallic Plethysmograph, by contrast to the other tests discussed in this article, is a physiological behavioral assessment device. The device measures:

". . . changes in blood flow through the penis [to] detect sexual arousal. By showing photographs of sexual stimuli to subjects, it is possible to determine their sexual preferences. Child molesters, for example, might be expected to exhibit sexual arousal at the sight of young children; homosexuals at the portrayal of someone of the same sex; and heterosexuals for members of the opposite sex. It is then possible to pinpoint features that subjects find most attractive, such as hair and eye color, height, bust measurements, or age." (Roedinger, supra, at 525.)

a. History And Background

The phallic plethysmograph was developed as an adjunct to biofeedback. Biofeedback measures a person's responses to stimuli and allows the person to learn to control his or her responses to those stimuli. Phallic plethysmography has been developing over the course of the last twenty years as a means of measuring male sexual response.

Speaking generically,

"Plethysmography may be defined as the measurement of changes in volume of a portion of the body. Since transient changes in the volume of most parts of the body are related to blood circulation, plethysmography can determine changes in blood volume in the part being examined." [K. Kedia, "Penile Plethysmography Useful in Diagnosis of Vasculogenic Impotence," Vol. 22, No. 3 Urology 235 (September, 1983).]

Phallic plethysmography has been used to diagnose and treat

impotence. In addition, in the legal forum, the Phallic Plethysmograph has been used to assess and treat sex offenders. Plethysmography combines the use of an electrocardiogram machine and some type of "strain gauge," or "constricting ring." [P. Malcolm, et. al., "Control of Penile Tumescence: The Effects of Arousal Level and Stimulus Content," Queen's University, Kingston Penitentiary Treatment Center, Canada Behavior Research and Therapy Vol. 23(3) 273-280 (1985).]

b. Construction of the Test

Although several types of methods have been conducted, typically, subjects undergo the following:

(1) The subject is seated on a recliner.
(2) The subject is isolated.
(3) The technician administering the test is either in the room, or capable of communicating with the subject by intercom. The technician might also be behind a two-way mirror to assure that the subject is watching the slides presented.
(4) The penis might be wrapped in something like Handiwrap to assure cleanliness.
(5) A plethysmographic cuff is wrapped around the base of the flaccid penis. (Note: the cuff is akin to the type of cuff typically used to measure blood pressure).
(6) The cuff is constricted (if a pneumatic cuff is used, the tightening is accomplished by injecting air into the cuff "producing an air pressure equal to mean arterial pressure plus one-third of the blood pressure); this procedure yields "reproducible" measurements. (Kedia, supra note 107, at 236.)
(7) Every action has an equal and opposite reaction. An increase or decrease in the size of the penis has a direct impact on the amount of air trapped within the pneumatic cuff.
(8) The subject is shown slides ranging from clothed males or females to nude subjects, to prepubescent or blank slides. The cuff measures the erectile response of the subject. The slides may be accompanied by an audiocassette, which vary the severity of aggression portrayed by the slide. (Id.)

Penile tumescence has been documented as the best measure of sexual arousal in males. [M. Zuckerman, "Physiological Measures of Sexual Arousal in the Human," In N.S. Greenfield and R.A. Sternback (Eds) Handbook of Psychophysiology 709-749 (New York: Holt Rinehart & Winston 1972).] Hence, the Phallic Plethysmograph is an attempt to document penile tumescence. (Malcolm, supra.)

The above-described method is only one of several available;

actual plethysmography measures the volume displaced by the enlarged penis. (Id.) Some methods employ circumferential measuring devices, including mercury-in-rubber strain gauges, [J. Bancroft, H. Jones & B. Pullman, "A Simple Transducer for Measuring Penile Erection, With Comments on Its Use in the Treatment of Sexual Disorders", 4 Behavior Research and Therapy, 239-241 (1966).] volumetric devices, [D. Wheeler and H.B. Rubin, "A Comparison of Volumetric and Circumferential Measures of Penile Erection", 16 Archives of Sexual Behavior, 289-299 (1987).] and metal-band strain gauges. [D.H. Barlow, R. Becker, H. Leitenberg & W.S. Agras "A Mechanical Strain Gauge for Recording Penile Circumference Change," 36 Journal of Applied Behavioral Analysis, 73-76 (1970).] There has not been a sufficient amount of comparative data compiled to determine if one method is more reliable than another. The gauges all measure increased gauge strain, directly corresponding to circumferential increases of the penis. Note: if ever there was an area calling for a case by case analysis, this is it. Obviously, the transformation from flaccid to erect will be vastly different between subjects. Many researchers therefore advocate conversion of circumferential gauge scores to percent full erection scores, thereby eliminating individual variances. [K. Freund, "A Laboratory Method of Diagnosing Predominance of Homo- and Hetero-Erotic Interest in the Male", 1 Behavior Research and Therapy, 85-93 (1963).]

c. Measuring Erectile Responses

Science has surely progressed since Mae West first asked "Is that a banana in your pocket, or are you just happy to see me?" The following methods are used in measuring responses to the Phallic Plethysmograph:

(1) Millimeter change in circumference from the flaccid state.

(2) Conversion of raw score into percentage score based on subject's measurement at full erection.

(3) Z-scores are determined by analyzing each subject's responses to all stimuli; mean and standard scores are calculated, then converted into "z" scores, which represent the mean in standardized deviation units. [V.L. Quinsey and G. Harris "Comparison of Two Methods of Scoring the Penile Circumference Response: Magnitude and Area", 7 Behavior Therapy, 702-04(1976).]

Researchers conflict over when to score the erectile response -- many opt for scoring penile tumescence in terms of peak response, while others measure the total curve of the response, which also measures latency. (Freund, supra.)

d. Use Of Microcomputers In Tumescence

Monitoring

Many innovators of Plethysmography advocate the use of "computer assisted tumescence monitoring systems." [W.R. Farrall and R.O. Card, "Advancements in Physiological Evaluation of Assessment and Treatment of the Sexual Aggressor", 528 Annals of the New York Academy of Science, 266 (1988).] The new affordability of personal computers means that every practitioner can have advanced technology at his or her fingertips. "Preliminary findings indicate that with improved procedures, a wealth of previously unseen data about the offender can be gathered. (Id.)

(1) Examples Of Deviant Audio Segments

Audio segments are used in the presentation of some Phallic Plethysmographs, followed by, or in conjunction with, slides "involving a wide variety of deviant sexual activities focused on male/female victims over a wide range, and an adult-heterosexual sequence." (Id. at 268.) While the authors refuse to present examples of deviant visual segments used in a Plethysmograph, following are two examples of deviant audio segments formulated by Robert D. Card, of the Clinic for Counseling and Psychotherapy, Inc., and William R. Farrall, of Farrall Instruments, Inc. (note: Farrall Instruments is the world's largest manufacturer of plethysmographs), who have developed a new, computer-controlled stimulus presentation, data collection and analysis system:

CHILD MOLEST, Female Victim 3-8 years Old. I really like little kids. They like me too. They always seem to want to climb on my knee or play horsie with me. I feel so good when I see them running to me when I go into the room. They want to climb on my lap and wriggle around and cuddle up to me. They're so soft and cute. I hope she wants to sit on my lap today. She's the cutest one. I feel so good when I'm there. I just want to touch her a bit. Maybe we can do it in the other room. It feels so good when she squirms around. Maybe she'll want to touch it. I could get her to do some things if we played some games. She seems to know all about it so it can't hurt her.

ADULT HETEROSEXUAL, Adult female. I'm really having a hard time concentrating on my work today, She's wearing that tight sweater again. If she comes over here once more and bends over my desk I'm not sure I can keep from grabbing them. I can almost feel the excitement if I could push my face between them. Get a hold of yourself! She looks almost as good from the back. Can't keep my eyes off her butt as she walks past. I can almost feel my hands running along that soft body. No wonder I'm so horny! I haven't had it for days. The last time we got it on was

wonderful! I've got to get some tonight. I can almost feel her squirming with excitement at the touch of my hands. (Id.)

e. Treatment Uses

For those who have been convicted of rape, child molestation and/or a related offense, the Phallic Plethysmograph can be used to treat and lessen deviant responses by juxtaposing arousing pictures with aversion therapy -- such as whiffs of ammonia. [R.E. Freeman-Longo and R.V. Wall, "Changing a Lifetime of Sexual Crime", Psychology Today, 58-64 (March, 1986).]

One practitioner in Dallas who administers the Phallic Plethysmograph for federal probation and state child protection agencies comments that the Phallic Plethysmograph is " . . . not the be-all, end-all, but it's the best tool we have now for evaluating these cases." [S. Crawford, "Device Used for Offenders", The Dallas Morning News, 41-42A (November 13, 1988).] Child protection agencies in both Dallas and Tarrant County have used the Phallic Plethysmograph "when trying to determine whether to allow a father accused of child molesting back into the home." (Id. at 42A.)

William Farrall, president of Farrall Instruments, Inc., the world's largest manufacturer of plethysmographs, has been quoted as saying:

We keep them as far away from law enforcement as possible. The quality of justice is less than perfect, and we don't want this used to force confessions out of people. (Id.)

While Mr. Farrall may have the best of intentions, deviant responses to the Phallic Plethysmographs have undoubtedly been used by prosecutors in even our fair-minded county to push an accused perpetrator over the brink toward a full confession.

Some researchers have noted the limitations and proper uses of the Phallic Plethysmograph:
". . . First the penile tumescence plethysmograph is not a sexual 'lie detector.' It will not tell whether a suspected offender has actually offended, nor will it tell if the offender is certain to offend again or not to offend. In other words, it cannot be used to search for probable offenders in the general public. However, properly obtained and interpreted, the penile evaluation can generally make it possible to determine the gender preference, age preference, and in many cases the type of sexual activity and interest to both an offender and a non-offender. This is obtained by noting the relative level of sexual response to various levels of stimulus materials." (Farrell & Card, supra note 120, at 262.)

Of extreme importance in assessing results of Phallic Plethysmograph is the realization that "sophisticated offenders can control their erections and often alter their response during assessment. Less knowledgeable offenders may learn from the assessment experience or are coached by those who have previously been assessed." (Id. at 264.)

f. Reliability/Validity

The physiological studies done from the 1960's to the present regarding the Phallic Plethysmograph may give the device an aura of credibility that it does not deserve. Almost all studies regarding the Plethysmograph conclude that there is a need for further research. Although the Plethysmograph has some value as a treatment device, questions arise as to the validity of the test outside of the context of the laboratory.

"The key question to the validity of penile erection measurements in diagnosing sexual deviants remains the comparison of those findings with the response of a normal population to the same erotic stimuli. Only then could the examiner assert with reasonable certainty that the subject who denied having a paraphiliac disorder but showed a positive arousal patters was in fact telling the truth." [S. Travin, K. Cullen and J.T. Melella, "The Use and Abuse of Erection Measurements: A Forensic Perspective," Vol. 16, No. 3, Bulletin of the American Academy of Psychiatry and the Law, 235, 241 (1988).]

In striving toward validation of Phallic Plethysmography, most researchers have compared and contrasted results of "deviant" subjects, such as the results of child molesters, to the results of other deviant subjects, such as rapists:

"Unfortunately, there have been few studies that have compared large normal subject groups with parapiliacs. The majority of comparative studies have tested intragroup differences, e.g., aggressive pedophiles versus nonaggressive pedophiles . . . there were no significant differences between the erection measurements or incarcerated rapists and incarcerated non-rapists." (Id.)

Current research certainly does not support the theory that results of Phallic Plethysmographs should be used as a means of proving guilt or innocence of a sex crime, or that a father should be deprived of visitation rights because of deviant scores regarding pedophilic slides. If used as part of a comprehensive analysis of a given subject, Phallic Plethysmography may have some modicum of credibility. Perhaps the most appropriate use of Phallic Plethysmography is in aversion therapy.

However, when used in the context of allegations of sexual molestation in a family law matter, the reliability and validity of the instrument should be seriously questioned.

g. Direct Examination Of The Expert

(1) Have the expert establish the subject's sexual history, and previous psychological and criminal records, if any.

(2) Have the expert establish who supervised the Phallic Plethysmograph, and the conditions under which the test was administered.

(3) Have the expert explain how measurements were taken - is the report based on raw scores or percentages/z-scores?

(4) Have the expert report whether the accused ever reached a full erection and if so, in response to what?

(5) Have your expert explain whether the penile tumescence scores were measured by a pneumatic cuff, a volumetric device, a mercury-strain gauge, etc. and whether the expert's opinion of reactions would be any different if any of the other means had been used.

(6) Have the expert report the time sequence between audiocassette deviant presentations and deviant slides, if both were used.

(7) Have the expert report if an audio presentation was used, whether explicit details of foreplay and were intercourse given, and if the details were violent.

(8) In the case of an accused rapist, have the expert differentiate reactions of a rapist from a non-rapist? from someone who engages in pedophilia? from a homosexual? from a non-offender? How?

(9) Also your expert should explain how he or she would recognize faking.

h. Cross Examination Of The Expert

(1) A subject may score a wide variety of false negatives and false positives on the Plethysmography - doesn't the range of possible inaccuracies make the test totally unreliable?

(2) Inquire as to whether the test taker was subjected to a blood exam prior to administration of the Plethysmographs? (Drugs and alcohol may have a direct impact on results. If a father in a custody dispute was accused of child molestation, and took a Valium prior to submitting to the test, the results could be skewed).

(3) Inquire as to whether the recording equipment used was standardized?

(4) Ask the expert: if another expert with his or her credentials was shown these test results, is there any guarantee that he or she would interpret the subject's responses in the same fashion you have?

(5) Inquire as to whether the expert would characterize visual materials as more stimulating than audio materials? (Card and Farrall

would assert that audio materials are more stimulating sexually than visual materials, due to the imagination factor).

(6) Ask the expert if it is true that sophisticated offenders can control their erections? Aren't sophisticated offenders capable of altering their responses during assessments?

(7) Ask the expert if there is any way to know if the test taker is thinking about the slide presently in front of them when an erection occurs, or whether the test taker may be daydreaming about a previous slide, or whether a prior slide has caused the subject to lapse into a fantasy.

(8) Inquire as to whether the subject was allowed to become detumescent between audio-visual presentations. If not, can the scores the subject received as the Plethysmograph progressed be considered accurate?

(9) Ask the expert if he or she considers the Plethysmograph as accurate as a lie detector? (If not, the Plethysmograph results have no more business being admitted than lie detector results).

(10) The expert should be questioned as to whether the test results be considered independent of a thorough review of the subject's sexual history. If so, it is time to impeach the expert!

(11) Question the expert as to whether he or she believes that the Plethysmograph results should be the sole basis for deciding whether or not the subject should be released from prison or from a treatment program? (If yes, you have an expert ready to be impeached).

5. Custody Quotient (Cq)

a. History and Background

The Custody Quotient (hereinafter referred to as the "CQ") is a new approach to assist psychologists in the evaluation of parents in child custody cases. The CQ was developed by Dr. Robert Gordon and Dr. Leon Peek, of Dallas, Texas, as a guide for explaining expert findings by report and in court. Further, the developers of the CQ see it as a guide for planning the remediation of a parent's shortcomings and how to use the parents strengths to the best of his or her ability. The authors of the CQ designed it because of a growing dissatisfaction with traditional psychological tests as they are applied to measuring good parenting.

b. Purpose

The primary purpose of the CQ is to assist Courts, attorneys and impartial third parties to resolve child custody disputes in a child's best interest. The CQ is also designed to have results that are relevant to determining sole and joint custody, including the assigning of parenting rights and duties as well as deciding issues of access. Additionally, the CQ results address the threshold issue of whether a material change of circumstance has occurred in a parent's life when a modification has been filed.

Another purpose of the CQ is to add to psychologist's fund of assessment procedures a device which is especially designed to address custody questions. The developers of the CQ felt that while there are a number of valid and reliable psychological procedures for assessing psychopathology, few have been researched for legal application and fewer still for custody issues.

c. Construction of the CQ

The CQ is a system of mapping judgments (similar to DSM III-R judgments) or the judgments of a diagnostic team about a person's capacity to be an effective parent. The CQ is not intended to direct the decisions of the court or jury regarding the custody of a child or children, but to assist in providing relevant information about the knowledge, attributes, and skills of adults as well as provide information relevant to the issue of modifying existing custody arrangements.

Additionally, the CQ leads to designing remedial programs for parents with low scores and therefore the CQ score and profile is not static by design. The parent may retake the CQ and document progress made on established goals every six months. Unlike most psychological

instruments, the person taking the test has an opportunity to study, learn, gain experience and then improve their score.

The CQ's Manual outlines an education-based therapy for acquiring new parenting skills and for retraining a parent with impoverished skills. In the absence of remediation or a significant event in the life of the parent, a CQ evaluation is considered current, by its authors, for up to six months. Following remedial therapy, retesting at intervals of three months is recommended. (R. Gordon and L. Peek, "The Custody Quotient: A Test of Effective Parenting to Assist with Custody Decisions", TRIAL INSTITUTE, The Texas Academy of Family Law Specialists, 1989.)

Gordon and Peek did something that to the authors' knowledge has never been done. Gordon and Peek researched Texas laws before allowing any item on the CQ. Therefore, there is legal authority for the relevance for each CQ item from case law, statute or by reasonable inference from the Uniform Marriage and Divorce Act.

Each parent response is graded with reference to written standards (similar to the Wechsler Comprehension Scale). In doing so, the examiner may consider other information acquired about the parent and child. The Frankness Scale on the CQ helps the examiner decide the degree of candor with which the parent disclosed information (similar to the L, K and F scales on the MMPI). The maximum time allowed for administering the standard CQ interview is two hours. (Id.)

Presently, the classification of CQ results are derived from the theoretical properties of the normal curve. For that reason, Gordon and Peek warn that until national norms are available, inferences about good parenting based on CQ results alone should be made with appropriate caution. (Id.)

The CQ is a set of ratings based on a composite of clinical procedures. The examiner may select from a variety of standard approaches practical for the particular case: the standard CQ interview, a clinical interview with history, objective and projective tests, a review of documents, observations of the parent-child interaction, a home study and collateral contacts.

The CQ avoids technical terms and is designed to be sex neutral. The comment sections in the CQ Manual points out cultural and intellectual differences among parties as they are known to affect psychological test conclusions and therefore the Manual points out that cultural differences must be kept in mind. (Id.)

Additionally, the attorney should keep in mind that the summary

CQ score describes the parent at a single point in time. But it considers parent performance over-time. A parent living in an intact family, who intends to divorce, will likely have a different profile six months after separation.

Examiner qualifications for administering the CQ is the American Psychological Association guidelines for test users. The CQ may also be administered by a trained assistant under qualified supervision. (Id.)

d. Definition of Good Parenting

The CQ Manual offers the following definition of good parenting:

"Those collections of attributes, skills and behaviors which adults rely on and use in raising the next generation. Good parenting occurs when adult practices lead the child to live independently and fulfill their biological, psychological and social potential. . .

The following psychological laws of good parenting are offered.

Law 1. Good parenting is designed to protect the child from harm during the child's vulnerable period of growth and maturation.

Law 2. Good parenting includes teaching the child skills for mastery over their environment so the child can live independently of their parents and after their parents have died.

Law 3. Good parenting creates an environment for the child conducive to the child fulfilling the biological, psychological and social potential.

Corollary 1. Those attributes, skills and behaviors of a parent which facilitate the operation of these laws are in the best interest of the child.

Corollary 2. Those attributes, skills and behaviors of a parent which interfere with the operation of these laws are not in the best interest of the child. [R. Gordon and L. Peek, supra.]

In CQ terminology, the adjectives "good", "effective" and "competent" are used interchangeably.

e. Development Of The CQ Scales

The CQ consists of ten clinical scales and one additional scale. The development of each scale was done in the following fashion:

(1) Expert Opinion

The expert opinion of the developers of the CQ was used to define a number of elements in the domain of good parenting. (Both Gordon and Peek are licensed psychologists with considerable experience in examining parents and children in custody disputes. In addition, Dr. Gordon is trained in law and
Professor Peek has taught psychometric research at the graduate level for many years).

(2) Theory of Child Psychology

Additional elements of the domain of parenting were gleaned from a review of theory and research in child psychology, child psychiatry and child development. The review also included literature in the fields of anthropology, philosophy, theology and comparative psychology. (For details see The Custody Quotient - Research - Manual, Chapter 2). (Id.)

(3) Home Studies

A systematic observation of several characteristics of the observable job of parenting was conducted based on video-taped home studies.

(4) Opinions of Parents and Children

Several studies were done of attitudes of parents and children toward good parenting and related issues. (For details, see The Custody Quotient - Research - Manual, Chapter 2).

(5) Opinions of Judges and Attorneys

A survey was conducted of district judges and family law specialists (that survey is in the process of being updated and presently approximately 450 district judges have responded nationwide).

(6) Based in Law

Gordon and Peek eliminated any item of parenting for which there was no authority in law to indicate its relevancy for custody decisions. The family code of various states, case law, the Uniform Marriage and Divorce Act and specifically Sec. 12 of The Texas Family Code were reviewed. (See Chapter 2 of CQ Manual for more detail).

(7) Categories

The refined items of good parenting were binned into categories of

parenting knowledge, attitudes, abilities, skills and behavior. A descriptive word or phrase was selected to represent each scale of items.

f. The CQ Scales

Each item of each scale receives equal weight, however, due to the study conducted of judges and attorneys, two scales receive greater weight.

(1) EMOTIONAL NEEDS SCALE (EN)

This scale consists of ten items that are directed toward how well the parent meets the emotional needs of the child. An example of the EN Scale is as follows:

"**EN 8** Parent willing to admit mistakes

2 parent has the maturity and insight to acknowledge errors in parenting or in decision making about child

1 parent seems reticent to examine their own errors or shortcomings in parenting or lacks the insight to do so; is overly defensive or superficial

0 parent seems incapable of recognizing their own shortcomings or errors; typical defense mechanisms are denial, rationalization and projection (blames child or the other parent or bad luck for things gone wrong)

_____

What decisions or actions concerning your child are you most proud of?

Are there some decisions or actions you've taken you now think were mistakes?

Comment No one is perfect. A mature parent is able to recognize their own shortcomings and make disclosures about errors in judgment. Emotional problems in children are sometimes caused by parents insisting that all of the failures and disappointments of the child are due solely to shortcomings in the child. The too obvious example is the parent who suggests to the child that the child is responsible for "Daddy or Mommy leaving home."

For superior ratings, look to parents who have made a point of

being clear with the child that the divorce was not the child's responsibility and that the decision of post divorce or modification plans is the job of the judge and not the child.

It is helpful to ask the child if their parents say they're sorry when they make a mistake. It is not proper for mental health professionals to encourage children to "inform" on their parents." (Id.)

(2) PHYSICAL NEEDS OF THE CHILD NOW AND IN THE FUTURE SCALE (PN)

This scale contains fourteen items that were designed to assess how well the parent takes care of the physical needs of the child. An example of the PN Scale is as follows:

"**PN 3** Provides healthy diet

2 parent reveals concern and awareness of nutrition; provides balanced diet; child occasionally helps prepare meals

1 parent knows and employs adequate concepts of nutrition and diet

0 parent usually takes child out to eat unhealthy fast food; orders out most nights; serves junk foods

Who prepares your child's meals at home?

How often does your child eat out? How often do you take fast food home?

In general, what are the most important aspects of a child's diet? (If vague) Could you please be more specific?

(For school age children) Does your child buy or take lunch to school? (If taken to school) Who prepares your child's lunch for school?

(If cooking is delegated) How did you pick the cook/housekeeper? How do you know they are doing a good job?" (Id.)

(3) NO EMOTIONAL OR PHYSICAL DANGER TO THE CHILD SCALE (ND)

The ND Scale like the ED Scale is given greater weight than the

other scales. There are eleven items on the ND Scale, an example of which follows:

"**ND 8 Child's Surroundings are Free From. . . Kidnapping/denying access**

2 parent does not have possession of child contrary to a court order; parent has not frustrated or denied other parent's access to child in person, by phone or mail contrary to court order

1 parent expresses willingness to kidnap child or to frustrate other parent's access to child contrary to a court order

0 parent has unlawful possession of child; parent has denied lawful access of other parent to child

---

(For custodial parent) Are there some circumstances under which you do not allow your child to see or talk to their other parent?

How do you keep the other parent informed of events, problems, programs and activities in your child's life?

(For non custodial parent) How often do you see your child and talk with your child by phone?

Is there a court order concerning the periods of time your child is in your home? (If yes) What does it provide?" (Id.)

(4) GOOD PARENTING: KNOWLEDGEABLE, ATTRIBUTES, SKILLS, ABILITIES, PARTICIPATION SCALE (GP)

This scale consists of thirteen items and is designed to measure the parenting skills of the parent. An example of the GP Scale follows:

"**GP 9** Parent is able to organize family affairs

2 parent coordinates the schedules of members of family to compliment each other; child gets places on time

1 parent tries to organize competing routines of family members; there is occasional tardiness or missed appointments; parent tends to rigidly adhere to schedules so that spontaneous opportunities for growth and learning are missed

0 parent seems unable to tolerate challenge to family organization; family events are happenstance; child often misses appointments or is

tardy

What is your child's usual daily schedule in your home, during the week and on the weekend?

What is the most difficult part of your child's schedule to keep on time? What do you do when this problem causes your child to run late?

Comment The examiner may wish to ask whether the parent keeps a calendar of the child's activities and those of other members of the family. The examiner may want to review the family calendar at a later session. Obviously organizing children's activities is more critical for school days than weekends or holidays. Conventional wisdom is that both chaos and overly compulsive organization have untoward results." (Id.)

(5) PARENT ASSISTANCE SCALE (PA)

This scale, consisting of seven items, is designed to determine how appropriately a parent is using help from others with their child. An example of the PA Scale follows:

"**PA 2** Family support (grandparents/other siblings)

2 grandparents/relatives live close by, are available, and desire to help with child; siblings are healthy and constructive

1 grandparents/relatives are only tangentially involved in child's life due to distance, lack of interest or poor health; are available only for celebrations of milestones in the child's life or for holidays; siblings are good models for child but are too busy to support parent

0 grandparents/relatives are uninvolved in child's life and are preoccupied with their own needs; grandparents are deceased; siblings have impoverished relationship with child

Are your parents living? (If yes) Where do they live?

When did they see your child last?

What activities did they do with the child on that occasion?

Are there other relatives who see your child from time to time? (If yes) Who are they? How often do all of you get together?

Does your child have a brother or sister, or a step brother or sister? What is the brother or sister like? How do they get along with your child? Give me an example of how the brother/sister helps you parent?

Give me an example of a way in which the brother/sister is not the best influence for your child?

Comment This item concerns the parents and relatives of the parent being interviewed and of the siblings living with the parent. As the life span of Americans increase and since half the American work force is comprised of women, grandparents and great grandparents play an increasingly important role in child's care.

Many grandparents have substantial contact with their grandchildren and provide emotional security and a sense of continuity. Although it is common for brothers and sisters and step siblings to deny liking each other or caring about one another, just the opposite is ordinarily true.

In evaluating the influence of an older sibling, common issues are: maturity, power, knowledge and loyalties, privacy and self-demarcation. If the older sibling has replaced the parent in terms of responsibility for the child, rate 'O' on this item. If there are no older siblings then omit this item." (Id.)

(6) PLANNING FOR THE CHILD SCALE (P)

Consisting of five items, the P scale is designed to measure how well the parent is planning for all of the child's needs. An example of the P Scale follows:

"**P 3** Planning for child's medical needs

2 parent knows name of physician for routine assistance; describes plan for helping child with special medical needs whether for disability or allergies; describes process of selecting and evaluating specialists and method of payment for future services

1 Parent knows something about child's medical history but is weak on details; cannot give names or selection process for specialists

0 Parent does not know physician; cannot relate how he/she wold select physician or gives poor process (e.g.,use whoever friend at work

uses); seems oblivious to the importance of routine medical care

Does your child have any special health problem? (FI yes) How are you going to help them with this problem?

Who does your child see to have annual physicals? How did you select the doctor?

(If none) How would you go about selecting a specialist if one was needed?" (Id.)

(7) HOME STABILITY SCALE (HS)

As the name of the scale suggests, this scale is designed to measure the stability of the parent's home. The HS scale consists of six items, an example of which follows:

"**HS 5** Stable Lifestyle

2 parent maintains consistent employment, stable relationships with friends; parent maintains predictable daily schedule or informs child of changes so child can reach parent if necessary

1 parent or situation has forced at least one major change during the past five years; parent is not reachable by child due to changing schedule except at certain times of day

0 parent makes arbitrary changes in pattern of living; parent frequently changes circle of friends in the home; child does not know visitors; frequent changes do not permit child to know general activities of parent

What are the most important changes you have made in your lifestyle since your divorce (or separation)? By lifestyle I mean such things as jobs, economic status, where you live, what you do for recreation, who you spend time with.

Comment This item involves an overall impression by the examiner of previous HS items." (Id.)

(8) ACTS/OMISSIONS SCALE (A/O)

The A/O Scale consists of eight items and is designed to measure parent misconduct. The following is an example on the A/O scale.

"**Acts/Omissions**

The personal standards of the examiner are not the guiding criteria for rating the parent on these items. The criteria is the contemporary community standard. Moreover, the examiner should refine rating on the community standard in terms of any known subcultural or ethnic group differences. Also, the examiner must guard against stereotypes based on gender when rating the parent for these items.

Methods of corroboration include interviewing the child and the other parent.

A/O 1 Parent misconduct

Misbehavior is any activity outside of the limits of what is accepted. This is different and more permissive than what is average or typical to a group or community. People will tolerate actions more atypical than the behavior they enact. Also, the misbehavior must have a direct, inferred or potential impact on the child in order to be covered by this item.

2 no sign, indication, suspicion or report of misbehavior

1 misbehavior has occurred but has been discontinued; parent taking steps to prevent repeat

0 ongoing misbehavior

Has anyone suggested that you have done something improper or illegal during the past ten years?

(If yes) What did they say? Is what they said true or partially true? What were the circumstances?

Was your child aware of the situation? (If so) Who told the child about it?" (Id.)

(9) VALUES SCALE (V)

This scale consists of six items and is designed to measure how well the parent is transmitting values which allows the child to distinguish right and wrong. The following is an example of the V scale:

"Comment This item refers to an attitude or belief that the parent holds. Most values are transmitted to the child implicitly or through role modeling. Clearly there are some matters about which a parent must have the last word such as the danger of traffic for young children or the use of drugs by teenagers, the school subjects children should take,or the sports they are active in.

V 2 Ethics of parent

2 parent words and actions illustrate moral and ethical values without contradictory actions

1 parent only occasionally voices appropriate values or behaves in a contradictory manner

0 parent shows or voices values which contradict those of the general community or denigrates the need for values

What are the most important things that you want your child to know about the difference between right and wrong?

How do you teach these principles to your child?" (Id.)

(10) JOINT CUSTODY SCALE (JC)

The JC Scale is designed to measure whether a parent is a candidate for joint managing conservatorship. The scale consists of ten items. The following is an example of the JC scale.

"**JC 8** Geographic proximity of homes

2 parent lives close by other parent; parent has realistic plans to provide travel between the homes (e.g., children are old enough to walk or bike between the two homes in normal weather)

1 parent voices intention to live near other parent; parent has weak

plans for travel between the homes by the child

0 geographical relationship of homes does not allow for safe, convenient or inexpensive travel

How far is it in minutes from your home to the other parent's home?

Do you or the other parent have immediate plans to move?

Would you be willing to move if it made it easier for your child to go from one home to the other? How about moving out of this city if it became necessary?

Comment When parents live remote from one another or live in neighborhoods with highly disparate socioeconomic status, joint custody arrangements usually fail. Joint custody arrangements are "usually" strained when one of the parents remarries." (Id.)

(11) THE FRANKNESS SCALE (FS)

The FS Scale is not one of the ten clinical scales but instead is an additional scale designed to measure whether a parent has a proclivity merely to answer the questions in a socially desirable fashion. The scale is similar to the "L" scale on the MMPI. When a parent gives too many answers to the frankness questions in a "socially acceptable" direction, motivation distortion or perceptual inaccuracy is inferred. The FS questions are dispersed throughout each of the other scales. An example of an item on the FS scale is as follows:

"**GP 14/F Frankness**

1 Parent frankly admits to occasions when he or she did poorly in handling child's misbehavior

0 Gives only superficial examples or speaks in generalities

Are there times when you have regretted disciplining your child as you did? For example, because you learned your understanding of a situation was wrong or your discipline was too harsh or too lenient?

Comment Appropriate discipline is a matter of parental discretion. Usually a parent has a preferred method for infants - facial expression, raising voice or momentary physical restraint; for the young child - spanking, withholding privileges, "time out", sending the child to their room or the "One minute scolding"; for teens - grounding, no telephone or television, or restricting driving for a period of time. The examiner should determine whether the parent's preferred method is consistent, predictable and produces the desired result.

As a method of corroboration, it is helpful to ask the child whether they think their parent's discipline is fair. There is a difference between a parent being authoritative and being authoritarian. Effective discipline does not include making the child think they are a bad human being. Nor does it involve embarrassing the child.

As is true of other CQ items, the examiner should consider the practices of the parent's subculture when rating the parent on this item.

The authors define abusive discipline as parent reactions which cause physical injury to the child or which diminish the child's self esteem. For extreme parent response, refer back to ND1 and ND2." (Id.)

g. The Custody Quotient Score

The CQ summary score is expressed as a standard deviation quotient score. This score is relatively easy to compare to other psychometric measures. The summary score lies on scale where the average for the general population is 90 to 110, "superior" begins at 120, and "borderline" begins at 79.

At this time, the "Classification System" is derived from the theoretical properties of the normal curve. The normal curve is the most common statistical distribution in nature, including established psychological phenomena.

The property of the normal curve important to the CQ scores is that a majority of persons are expected to fall within plus or minus one standard deviation, closely approximating the average of the general population. A CQ score of 85 to 115 will contain about 68% of the general population. A CQ score of 70 to 130 will contain about 96% of the general population. (Id. at 9.)

h. The CQ Classification System

CQ/JQ Parent Grand Score	Competency Range	% of Parents in Range
130	Very Superior	2.2
120-129	Superior	6.7
110-119	High Average	16.1
90-109	Average	50.0
80-89	Low Average	16.1
70-79	Borderline	6.7
69 and below	Dangerous	2.2

(Id.)

i. Validity And Reliability

It must be remembered that the CQ is presently still in the research stage and therefore the studies of validity are presently underway. The CQ should therefore be treated with that in mind. However, unlike The Bricklin Perceptual Scales (BPS) and The Sex Abuse Legitimacy Scales (SALS), studies are being conducted to determine validity and reliability and the CQ is consistently being updated as new data comes in.

Specifications for the CQ call for a mixture of content, construct and criterion validity.

(1) Content Validity

The developers of the CQ identified a domain of knowledge, attributes, attitudes, skills and behaviors of adults which are relevant to the tasks of measuring effective parenting. The items themselves grew out of a review of interdisciplinary research and through the collective experience of the authors. Their relevance was corroborated through the understandings of children, parents, judges, attorneys and the CQ Panel on Development. (Id.)

(2) Construct Validity

Definitions and laws of good parenting were offered to give specificity to the domain of good parenting. A content analysis of the understandings of parents, children, attorneys and judges was made to define the parameters and boundaries of good parenting the CQ instrument intended to address. Direct observation of parents and children in a professional office environment and through videotaped home studies was also relied upon. Construct validity makes inferences possible from CQ

scores to the domain of good parenting. (Id.)

(3) Criterion Validity

At this point it seems unlikely that winning a custody lawsuit will emerge as the generally accepted criterion against which to validate psychological tests. There are no alternative measurements which could be used as criteria. (Id.)

(4) Classification

The classification system for CQ ranges of effective parenting were derived from the mathematical principles of the theoretical normal curve. The upper range was emphasized in order to enhance fairness during the CQ development period. This approach parallels the construction of classification systems in natural science. When normative studies are complete, the basis of classification will shift to empirical data. (Id.)

(5) Item Format

Items and standard interview questions were reviewed by three persons with known writing skills for appropriate grammar, the elimination of bias in syntax and construction flaws such as ambiguity. Questions were refined following the administration of the first 100 tests based on the comments of test takers. (Id. at 10)

(6) External Validity

External validity was subject to a General System Theory Approach (GSTA). The cross-examination technique was employed. Sample CQ results were presented and attacked in the mock courtroom of the Wilmington Institute by volunteering members of the CQ task force. Basic rules of evidence were observed.

The General Systems Theory Approach postulates that traditional methods of establishing validity do not consider the Gestalt of an instrument that is intended to measure a complex system of behavior such as good parenting. (Id.)

(7) Reliability

The reliability of the CQ will be established by evaluating the degree to which two or more expert examiners will give a particular parent the same CQ score and the degree of internal consistency that exists among CQ items and between CQ scales. A pilot study of interrater

reliability found high rating reliability between two experienced raters. (Id.)

(8) Research Status

The CQ is both a clinical instrument, which summarizes the ratings of experts along standard criteria, and a psychometric test. At this state of development, it is a "clinical instrument".

Clinical instruments serve as guides to experienced examiners and provide for a consistent way to record the opinion of the experts. Clinical instruments are "wide band" in that many attributes of the person are weighed and assessed when the examiner makes the ratings. A typical psychological test by contrast, scores attributes in a more or less mechanical way. (Id.)

(9) Standards and Guidelines

The CQ is being developed with reference to the standards for psychological tests of the American Psychological Association. Application is being developed with reference to constitution precepts of due process.

At this time the CQ has the specificity of administration typical of some well-designed psychological tests. Research is underway that will determine whether the scoring of the CQ is also sufficiently specific. This distinction has to do with the type of information assessed and the nature of the user. (Id.)

j. Direct Examination Of The CQ Expert

(1) Establish with the expert that while the CQ is a research device, it represents probably the best range of information that should be assessed in a custody battle.

(2) Have the expert explain whether the results to the CQ represent new or old behavior for the subject, and whether the primacy of the behavior should have an impact on test results.

(3) The expert should report on whether he or she has based his or her opinion exclusively on the CQ, or whether results to other tests and methods have also been addressed in reaching whatever recommendation is made to the Court. (If the CQ is given appropriately, many other factors are taken into consideration by the expert).

(4) Establish the expert's knowledge of the ongoing validation process being assembled regarding the CQ.

(5) Establish that because the CQ is standardized, the expert has guideline to compare two different parents.

(6) Establish that the CQ is a comprehensive interview system.

(7) It would be helpful to the fact finder to know some of the items on the CQ. Any judge or jury will see the relevancy of the items on the CQ especially as compared to "evil spirits possess me at times" from the MMPI.

(8) Establish with the expert that no other psychological instrument of assessment has justification for each item in the law.

k. Cross-Examination Of The CQ Expert

(1) When cross examining the expert who has given the CQ, and it did not come out favorable for your client, establish that the CQ is not a test.

(2) The expert should be cross-examined about the fact that the CQ is not yet supported by normative data.

(3) Get the expert to admit that the CQ's rating system is presently unsupported by empirical data.

(4) The expert should also admit that at present the CQ lacks external validity.

(5) If the expert is using the CQ to say who should have custody, then get the expert to admit that even Gordon and Peek say that that is not the purpose of the CQ.

(6) If only one parent took the CQ, there is no means of comparison, and the expert should admit that this lessens the value of the CQ.

5. Uniform Child Custody Evaluation System (UCCES)

Since no existing psychological test has emerged as **the** test to be utilized in custody evaluations, and because the existing tests are so dependent on the skill of the clinician in making reasonable interpretations in custody determinations, Dr. Harry Munsinger and Dr. Kevin Karlson developed a uniform interview procedure for **all** mental health professionals, called the Uniform Child Custody Evaluation System (UCCES). Initially, it is being published as an interview and decision-making guide while validation data is collected by the publisher, Psychological Assessment Resources. Additionally, the American Psychological Association's Committee on Professional Practices and Standards is currently in the process of developing standards and guidelines for administering child custody evaluation tests such as those previously discussed.

The hope is that in a few years, the UCCES will prove to be valid and reliable standard assessment procedure for use in custody evaluations

to that the primary focus of custody litigation can once again be the children and their parents and not the tests and who gave them. This test should be available by the end of this year.

V. CONCLUSION

A. General Considerations When an Expert has Psychological Testing

The first thing an attorney should be cognizant of is that, except with rare exception like the Bricklin Perceptual Scales, CQ, SALS, and UCCES, psychological tests were not designed to assess legal issues. Psychological tests are designed as a means to help the expert evaluate a patient, not to be taken into the courtroom and used as an end in themselves.

Putting aside the validity of each individual psychological test, the attorney should evaluate whether the clinician has used the test beyond its limits. Psychological tests, as stated several times in this article, are merely tools for the clinician, and as such, their usefulness is limited by the experience, expertise, and integrity of the clinician and of the purpose of the test. If the developer of the test or the clinician fail to understand the following concepts, then that failure to understand can lead to abuses of the tests and the test data:

1. The psychological tests which are used most in the United States are based on norms consisting of middle class Caucasian individuals. Therefore, the test may be valid when given to middle class Caucasians but invalid when given to persons outside of those characteristics. (Id. at 62.)

2. All psychological tests come with standard administrative instructions. The validity of each test is dependent on there being no significant deviation from those instructions. The effects of the deviation of the test results may be difficult to ascertain. (Id. at 62.)

3. How well an individual performs on a given test is often determined by the setting in which it is given. Individuals with similar psychological characteristics or intelligence may have quite different test results, dependent upon whether the test is given in a noisy waiting room of a clinician's office or in a quiet room off by itself. (Id. at 62.)

B. Standards and Guidelines for Child Custody Evaluations in Divorce Proceedings

Attorneys and psychologists alike should be aware of national

child custody evaluation guidelines promulgated by the American Psychological Association which are attached as Appendix "A." The guidelines focus on the psychologist's role in the divorce proceeding and provide guidance and standards for the psychologist in preparing the child custody evaluation. The guidelines address three major areas:

1. The purpose of a child custody evaluation;
2. Preparing for a child custody evaluation; and,
3. Conducting a child custody evaluation.

The major principle set out by the guidelines reiterates the precept found in nearly every state's family code or statutes - the child's best interests and well-being are paramount. Applying this rule to child custody evaluations, the guidelines provide that the primary purpose of a child custody evaluation is to assess the best psychological interests of the child. To best achieve this goal, the focus of the evaluation should be on parenting capacity, the psychological and developmental needs of the child, and the resulting fit.

In preparing for a child custody evaluation, the guidelines define the role of the psychologist as a objective, impartial, professional expert who does not attempt to act as a judge. Moreover, the psychologist should have current and specialized training or knowledge in order to competently undertake a child custody evaluation. The psychologist should also be aware of any personal and/or social biases that he might have which could affect his opinions in the evaluation process. Additionally, the psychologist should avoid any counseling or therapeutic roles with any of the parties and confine himself solely to his defined task.

In conducting a child custody evaluation, the psychologist should limit the scope of the evaluation to issues raised by the referring person or court, and obtain informed consent from all appropriate participants. Informed consent includes informing the participants as to the limits of confidentiality and disclosure often required in custody proceedings. The psychologist should also use every relevant and helpful available resource, while guarding against inappropriately interpreting or assessing the data. Finally, the psychologist should maintain detailed records in accord with the relevant statutory guidelines.

While many of the individual guidelines state what appears obvious to litigation-wise psychologists, they serve an important purpose because they establish the foundation and limits for a proper custody evaluation. This provides a reference point for the family law attorney who is overwhelmed and confused by the unfamiliar world of child psychology. From the perspective of the psychologist, the guidelines establish coherent standards for divorce and custody related practice.

C. Final Thoughts

It seems clear that opinions based on psychological tests are, at best, controversial and sometimes may even be unconstitutional. This does not mean that psychological tests are not effective tools for the clinician in a clinical situation; however, it is time that the legal community questions the use of the presently established psychological tests in the courtroom and, most specifically, in family law courts. There are no studies which suggest that, of the standard tests which are administered to custody participants, there is any validity in using those tests to assess parenting skills. Furthermore, there is an astounding amount of evidence showing that the standard tests do not "meet the criteria of established scientific principles or general acceptance by the scientific community." (Ziskin, supra note 2, at 250.)

At the very least, the attorney should use this article, and the sources cited herein, to become acquainted with information beyond our normal expertise. Without such information, an attorney cannot effectively perform direct or cross-examination of an expert who has administered any of these tests.