TEST-RETEST RELIABILITY ON THE REVISED

CONNERS' PARENT RATING SCALE

THESIS

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

For the Degree of

MASTER OF SCIENCE

By

Debra Gomez, B.S.
Denton, Texas
May, 1998

The purpose of this study was to determine the test-retest reliability of the Revised Conners' Parent Rating Scale. The Conners' Parent Rating Scale (CPRS-48) was administered to 59 elementary school children between the ages of 5 and 10 years. After a period of two weeks, the same children were re-tested with the CPRS-48. A Pearson Product Moment Correlation was calculated for each of the CPRS-48 scales. The current study produced adequate test-retest correlation coefficients ranging from .92 to .75. This study found no significant differences in the consistency of scores overtime and the gender of the subject, gender of the rater, or age of the child. The results of this research lend support to the integrity of the test-retest reliability of the CPRS-48. The need for further psychometric studies on the Conners' Scales is noted.
TEST-RETEST RELIABILITY ON THE REVISED

CONNERS' PARENT RATING SCALE

Thesis

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

For the Degree of

MASTER OF SCIENCE

By

Debra Gomez, B.S.
Denton, Texas
May, 1998
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>1. LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>Primary Types of Assessments</td>
<td></td>
</tr>
<tr>
<td>Checklists and Behavior Rating Scales</td>
<td></td>
</tr>
<tr>
<td>Reliability and Validity of Checklists</td>
<td></td>
</tr>
<tr>
<td>The Conners' Behavior Rating Scales</td>
<td></td>
</tr>
<tr>
<td>Reliability and Validity of the Revised Conners' Scales</td>
<td></td>
</tr>
<tr>
<td>Statement and Purpose of Hypothesis</td>
<td></td>
</tr>
<tr>
<td>2. METHOD</td>
<td>21</td>
</tr>
<tr>
<td>3. RESULTS</td>
<td>25</td>
</tr>
<tr>
<td>4. DISCUSSION AND IMPLICATIONS</td>
<td>29</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>33</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>36</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table

1. Reliability Coefficients for the Scales of the CPRS-48 ......................... 26
2. Hotelling T Square Test for Gender Difference of Subjects .................. 27
3. Hotelling T Square Test for Gender Difference of Raters .................... 27
4. Hotelling T Square Test for Age Difference ..................................... 28
Primary Types of Assessments

Achievement

In the school systems children are often assessed for academic achievement, intellectual ability, and under special circumstances for behavioral and personality problems. The results of these assessments are used to make both clinical and nonclinical decisions about children. Kaplan and Saccuzzo (1982, p. 4) describe a psychological test as, "a device for measuring characteristics of a human being that pertain to behavior."

Anastasi (1958, p. 23) states, "a psychological test is essentially an objective and standardized measure of a sample of behavior."

The focal point of the present research is child assessment and evaluation in the schools; more precisely, the reliability of behavioral or psychobehavioral assessment of school-age children. Schools are among the most substantial users of tests because of their application in placement decisions about children. The primary types of tests used in the schools are: achievement, aptitude, intelligence, personality, and behavioral.

Wodrich and Kush (1990, p. 129) state, "Achievement tests are designed to measure how well academic skills have been learned." Achievement tests are used for
three main purposes: to survey the performance level of a group, to diagnose an individual's strengths and weaknesses, and to gauge a student's readiness for the next stage in training (Goleman, Engen, & Davids, 1982). Achievement tests stress mastery of factual information (Sattler, 1988). Standardized achievement tests are typically capable of being examined for objective evidence of reliability and validity. Unfortunately, these tests are generally not able to identify precise academic deficits. They tend to be heavily dependent on formal learning acquired in school or at home, to be more culture-bound, and to sample more specific skills than do intelligence tests (Kaplan & Saccuzzo, 1982). A commonly used achievement test is the Wide Range Achievement Test - Revised (WRAT-R) (Jastak & Wilkinson, 1984). In the elementary schools the WRAT provides an estimate of grade level functioning in reading, spelling, and arithmetic.

Aptitude

Aptitude refers to the potential for learning a specific skill. Klein (1977, p. 96) believes "an aptitude test measures characteristics that are supposed to be predictive of future performance." Achievement is commonly referred to as previous learning, while aptitude is the potential for learning a specific skill. Aptitude tests are usually utilized for vocational and educational guidance. The technical difference between an aptitude test and an achievement test is easier to state in theory than in terms of test construction. Ordinarily, school learning experiences will affect those skills sampled by aptitude tests to some extent. Anastasi (1988, p. 412) states, "The aptitude tests measure the effects of learning under relatively uncontrolled and unknown conditions." An example of an
aptitude test is the Strong Vocational Interest Blank (Hansen & Campbell, 1985). This particular test is used to assist in determining vocational interests and offering guidance.

**Intelligence**

There are a variety of definitions of intelligence. Traditionally, intelligence been defined as a person's learning potential independent of prior learning (Kaplan & Saccuzzo, 1982). Binet defines intelligence as "one's capacity to (a) find and maintain a definite purpose or direction, (b) adjust strategy, if necessary, to achieve that purpose, and (c) evaluate or criticize so necessary adjustment could be made" (Kaplan & Saccuzzo, 1982, p. 227). Wechsler's (1958) definition states that intelligence is composed of qualitatively different abilities. He adds that intelligence is not the mere sum of abilities, because intelligent behavior is also affected by the way abilities are combined and by the individual's drive and incentive. Kline (1991, p. 143) specifically states, "The factor analysis of tests of ability (speaking here of intelligence tests) unequivocally yields two g factors labeled by Cattell (1971) as fluid and crystallized intelligence." Kline feels that fluid intelligence is our basic reasoning ability, dependent on the neural efficiency of the brain. Crystallized ability is the set of skills, valued by our culture, in which this ability is invested.

Despite all the advances in intelligence testing, there is still no absolute definition of intelligence. The Kaufman Adolescent and Adult Intelligence Test (Kaufman & Kaufman, 1993) is based on an integration of three developmental models of intelligence: Horn and Cattell's theory of fluid and crystallized intelligence(1966, 1967), Piaget's highest stage of cognitive development, the stage of formal operations (Inhelder & Piaget,
1958; Piaget, 1972); and the Luria-Golden definition of planning ability (Golden, 1981; Luria, 1980). Kaufman (1990, p. 25) states, "An intelligence test is a static set of stimuli designed to yield scores derived from a single model of intelligence--the test author's. Because the test designers begin with different ideas, test focus and items can be very different."

Kaufman (1979), claims that the intelligence test is really a kind of achievement test; not the same type of achievement test as reading or science, but a measure of past accomplishments that is predictive of success in traditional school subjects. He argues that when intelligence tests are regarded as measures of prior learning, the issue of heredity versus environment becomes irrelevant. He believes that since learning occurs within a culture, intelligence tests obviously must be considered to be culture-loaded, a concept that is different from culture-biased. Most general criticism of intelligence tests focuses on issues of cultural fairness or test bias. Two popular intelligence tests include the Wechsler Intelligence Scale for Children (WISC-III) (Wechsler, 1991) and the Stanford-Binet Intelligence Scale: fourth edition (Thorndike, Hagen, & Sattler, 1986).

**Personality**

Personality tests are instruments for the measurement of emotional, motivational, interpersonal, and attitudinal characteristics, as distinguished from abilities (Anastasi, 1988). Goleman, Engen, and Davids (1982) note that there are two main ways to measure personality. The direct method, objective testing, includes the use of questionnaires and self-rating scales. The indirect method uses projective techniques to probe unconscious factors in personality. Unstructured, ambiguous material such as perceptions of inkblots
are used to observe how one responds. The person's perceptions and interpretations reveal unconscious motivations and hidden personality structures.

**Psychodynamic personality assessment.** Within the field of personality testing there are two theoretical approaches, the behavioral and the psychodynamic. Psychodynamic assessments define overt manifestations of psychological behavior as symptoms of an underlying trauma. Uncovering or discovering this "cause" is the target of psychodynamic assessment and therapeutic procedures. Behavioral assessment focuses on the symptom itself as the problem rather than a sign of more basic underlying conflicts.

A large number of psychodynamic personality tests are classified as projective. The stimulus is presented in unstructured form allowing the client to "project" his or her own pathology (or lack of pathology) onto the test item, revealing possible inner psychic difficulties (Kaplan & Saccuzzo, 1982). An example of a projective test is the Rorschach Inkblot Test (Rorschach, 1942), the stimulus being an inkblot. Another projective test is the Thematic Apperception Test (TAT) (Murray et al., 1938). In this test, the stimulus consists of ambiguous pictures depicting a variety of situations and scenes.

Projective tests have been under much criticism concerning their usefulness in making diagnostic decisions. Wodrich and Kush (1990, p. 177) believe that "projective measures typically lack standard administration and scoring rules and certainly lack the quantifiable, norm-referenced results that their objective counterparts have." Anastasi (1988, p. 621) states that "besides their questionable theoretical rationale, many projective techniques are clearly found wanting when evaluated in accordance with test standards." Projective measures rely on the experience and clinical insights of the examiner. Because
of potential differences among test interpreters, projectives can significantly affect the relevancy of diagnosis, treatment planning, and educational service.

The more direct method of personality assessment uses objective evaluation of the individual's own behavior based on self-report, or through information provided by a knowledgeable informant. In sharp contrast to projective measures, objective measures usually have standard administration and scoring rules, and results are norm-referenced. Objective techniques commonly use checklists or true-or-false items. These techniques may provide a descriptive statement requiring the client to choose among a number alternative responses. It should be noted that some tests that follow the checklist or true/false format are interpreted under the psychodynamic model having the results linked to underlying inner-psychic difficulties. An example of this type of test is the Minnesota Multi-Phasic Inventory- 2nd Edition (Hathaway & McKinley, 1990).

Difficulties with the objective form of assessment may include respondents faking answers, respondent bias, misinterpretations of the questions, misconceptions when reports are filled out by observers, as well as any problems with the respondent understanding the question (Ollendick & Hersen, 1984). Rosenthal (1966) reports evidence that the observer's recordings (Kennedy & Uphoff, 1939; Scheffield et al., 1952; Rosenthal et al., 1964), interpretations (Rapp, 1965; Smith & Hyman, 1950), data fabrication (Azrin et al., 1961; Rosenthal & Lawson, 1964), and even the behavior of the person being observed (Rosenthal & Fode, 1963; Masling, 1965) are likely to be influenced by the observer's expectations.
**Behavioral personality assessment.** In contrast to the psychodynamic model, the behavior model assumes that overt behaviors are the focus of behavioral assessment and change. Silva (1993) listed the following definition by Kanfer (1979, p. 37): "Behavioral assessment is defined...as assessment done by researchers, clinicians, and other practitioners whose goal is to measure aspects of people's overt, publicly observable behavior, rather than covert, private events or constructs like thoughts, attitudes, or personality attributes." In dealing specifically with the recognition and evaluation of child behavior disorders, assessment has fundamentally been difficult. To the behavioral psychologist, the behavior is the problem; therefore assessment must provide the ability to sample behavior from a number of different contexts in the child's life. Sattler (1988) suggests the appropriate assessment of children with behavior disorders requires an extensive, multimethod approach. He feels assessment must occur across diverse situations, such as home, school, clinic, and community. He also notes that cross-situational assessment is complicated by the need to involve a number of different informants and to use multiple methods of collecting information.

**Checklists and Behavior Rating Scales**

**Advantages**

Recently in the area of child behavior disorder assessment, checklists and/or behavior rating scales have become very popular. Most checklists or rating scales for young children are completed by knowledgeable informants, usually parents or caregivers. The checklist allows a study of how each informant views the child's behavior and how one informant's views compares with another's. Since most children have trouble
expressing themselves verbally, heavy reliance is placed on information given by significant adults. This approach has found favor among behavioral psychologists, because it helps to define behavior categories as stated above. Checklists provide a direct assessment approach as opposed to the more indirect methods of projective techniques.

Behavior checklists have many advantages. These instruments are used primarily to generate information that is helpful in making classification, training, and intervention decisions. They are used with a wide variety of children with or in suspicion of behavior disorders. They provide information about a child or group of children which complements information obtained from other assessment procedures (Sattler, 1988). The checklists (or behavior rating scales) are (a) usually quick, simple, and inexpensive; (b) sensitive to treatment effects, like drug effects versus placebo effects; (c) measurable in the child's natural environment (not in a clinical one where observed behaviors may be unrepresentative); and (d) made by one who is familiar with the child (Zentall & Barack, 1979).

Checklists can be used as routine screening devices in various settings including: schools, outpatient clinics, inpatient clinics, residential treatment centers, child protective services, special education and regular classrooms, juvenile detention centers, and private practice (Conners, 1990). Behavior checklist ratings reflect a complex character interaction between the scale or checklist used (scale values, wording of items, content of items, standardization sample); the child (age, sex, type of disturbance); the informant (expectancies, recall ability, response bias, comprehension of items); the examiner (sex, ethnicity, traits); the setting (school, home, playground, hospital, prison); and the reasons
for the evaluation (screening, diagnosis, placement, program evaluation) (Sattler, 1988).

Aside from ease of administration, low cost, and quick interpretation, checklist ratings provide one more piece in the total diagnostic package without major time consumption. The Conners' Behavior Rating Scales (Conners, 1990), The Child Behavior Checklist (Achenbach & Edelbrock, 1986), and ADDH Comprehensive Teacher's Rating Scale (Ullmann, Sleator, & Sprague, 1991) are just a few examples.

**Integrity**

In order to evaluate the integrity of any checklist, one must review its reliability and validity. This is typically done by investigating: (a) the characteristics of its standardization sample, (b) the types and quality of its reliability, (c) the reliability and validity of comparable tests, (d) the scoring procedures, (e) the method of administration, (f) the limitations, and (g) the strengths (Kaplan & Saccuzzo, 1982). Whether testing is for adults or children, individuals or groups, the same basic test characteristics should be examined before the decision is made as to what test will be best for the assessment needed. The goal of any psychological assessment is to make valid decisions based on test data.

Some specific characteristics of well-normed behavioral checklists are they: (a) identify the examinee's behavioral strengths and weaknesses; (b) provide an objective basis for evaluating the examinee's results or progress in an intervention program; (c) permit comparison of the examinee's behavior in different situations (like home or school); (d) ease comparison of information from different informants; (e) provide a standardized way
of reporting information within and between organizations; and (f) stimulate new research and intervention programs (Sattler, 1988).

**Disadvantages**

Although behavior checklists can provide valuable assessment information, there are several factors which may challenge their validity. As mentioned above, ratings on any checklist reflect the interaction of the characteristics of not only the checklist and the child, but also the informant or rater, examiner, setting, and reasons for the evaluation. The credibility of the informant is a critical factor in evaluating information obtained. The ratings of multiple informants should be compared and contrasted. An informant's judgment may be affected by familiarity with the child, sensitivity and tolerance for behavior problems, personality, and expectations. Some raters may be influenced by the child's race, socioeconomic status, appearance, or psychopathology, any of which would make their ratings invalid. Observations may not provide congruent data (Sattler, 1988). Lastly, as is true for any form of assessment, if the test cannot demonstrate strong reliability and validity, it is of no use to the evaluator. For the purpose of this study, the primary focus will be on issues of reliability and validity.

**Reliability and Validity of Checklists**

**Reliability**

Kaplan and Saccuzzo (1982, p. 114) state "reliability is one of the basic foundations of behavioral research. If a test is not reliable, it will not be possible to demonstrate that it has any meaning." Conners (1990, p. 39) states, "To say that a scale, or instrument, is valid rests upon the weight of accumulated evidence from a number of"
validity using various methodologies." Anastasi (1988, p. 27) states, "Reliability refers to the consistency of scores obtained by the same person when reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions."

Reliability is one of the foundations of behavioral research covering several aspects of score consistency. Test reliability indicates the extent to which individual differences in test scores are attributed to "true" differences in the characteristics under consideration and the extent to which they are attributable to chance errors. Kaplan and Saccuzzo (1982), suggest that reliability estimates in the range of .70 and .80 are good enough for most purposes in research. The following is a review of the various forms of reliability. The first type is the reliability of the ratings given by two or more raters: This is called inter-rater reliability (Morris, Fitz-Gibbon, & Lindheim, 1987). Zentall & Barack (1979) claim interrater reliability, involving child observation, will be most accurate when the context, familiarization time with involved children, and the role of the observer (e.g., teacher, parent, counselor) are held constant. Henderson, Morris, and Fitz-Gibbon (1987) state, "the problem of inter-rater consistency is diminished when all recorders have learned to use the same set of rules for interpreting the behavior they witness."

By using alternate forms of the same test, the same person can be tested with one form on the first occasion and with another equivalent form on the second. Alternate form reliability should always be stated with the length of interval between test administrations. If the two forms are administered in two immediate successions, the resulting correlation would show reliability across forms, not situations (internal consistency). Alternate form
reliability has certain limitations; if the behavior under consideration is subject to a large practice effect, the use of alternate forms will not eliminate such an effect. Another concern is the degree to which the nature of the test will change with repetition. For example, an item involving the same principle can be solved readily by most examinees once they have worked out the solution to the first. Finally, alternate forms are unavailable for many tests, because of the difficulties of constructing truly equivalent forms (Anastasi, 1988).

Internal consistency based on either subtests or items is basically a measure of homogeneity; it helps characterize the behavior domain or trait sampled by the test. Internal consistency can be found using alternate-form or split-half reliability, which involves dividing a test into two equal halves that are scored separately. The results are then compared with each other. Though it may appear convenient, this method can cause problems when items on the second half of the test are more difficult than items on the first half (Kaplan & Saccuzzo, 1982). Split-half reliability estimates are not appropriate for timed tests and do not take into account changes over time. Generally, the size of the internal consistency coefficient is increased with greater test length (Sattler, 1988).

One of the most common ways to examine reliability is to repeat the identical test on a second occasion. Test-retest reliability shows the extent to which test scores can be generalized over different occasions. The higher the reliability, the less susceptible the scores are to random daily changes in the testing environment or in the condition of the test takers. Whenever reliability is reported, the interval between measures should always be specified. Short-range random fluctuations that take place during intervals ranging
from a few hours to a few months are generally included under the error of variance of the test score. In checking this type of test reliability, the main effort is to keep the interval short. In testing young children, the period of time between tests should be even shorter than for older persons, since at early ages progressive development changes are observable over a month or even less. For any type of person, the interval between retests should rarely exceed six months (Anastasi, 1988).

Although apparently simple and straightforward, the test-retest technique can present difficulties for certain tests. According to Anastasi (1988), this type of analysis is only valuable when measuring "traits" or characteristics of individuals that are not believed to change over time (stable traits). There is the possibility of a carry-over effect. This effect takes place when the first testing session influences scores on the second one, when test takers remember their answers from the first administration. When there are carry-over effects, the test-retest correlation usually will overestimate true reliability.

As mentioned earlier, a similar problem known as practice effect relates to skills that improve with practice. When given a test the second time, test takers score better because they have sharpened their skills by taking the test the first time. As a result, scores will be higher on the second administration. Because of these problems, the time interval between testing sessions must be chosen carefully. If the administrations are too close in time, there is a greater risk of carry-over and practice effects. However, as the time interval increases, many other factors intervene and are possible explanations of the differences between scores on the two testing sessions (Kaplan & Saccuzzo, 1982).
Validity

Along with examining reliability, a test must prove strong validity. Sattler states (1988, p. 30), "The validity of a test refers to the extent to which a test measures what it is supposed to measure, and therefore the appropriateness with which inferences can be made on the basis of the test results." Anastasi (1988, p. 28) maintains, "The validity of a test concerns what the test measures and how well it does so." Kaplan and Saccuzzo (1982, p. 117) state that "validity can be defined as the agreement between a test score or measure and the quality it is believed to measure."

Following are the different types of validity:

1. Face validity is the mere appearance that a measure has validity. It is often said that a test has face validity if the items are reasonably related to the perceived purpose of the test (Kaplan & Saccuzzo, 1982).

2. Content validity refers to whether the test items are representative of the domain that the test purports (Sattler, 1988). This form of validity will not be covered in this research due to the difficulties in separating it from other types of validity.

3. Criterion-related validation indicates the effectiveness of a test in predicting an individual's performance in specified activities (Anastasi, 1988). Performance on the test is checked against a criterion (a direct and independent measure of which that test was designed to predict). A form of criterion validity is predictive validity, this refers to the correlation between test scores and performance on a relevant criterion where there is a time interval between the test administration and performance on the measure (Sattler,
1988). In other words, is the score obtained on the test an accurate predictor of future performance on the criterion?

The second form of criterion validity is construct validity. Construct validity refers to how well the instrument measures what it claims to (Morris, Fitz-Gibbon, & Lindheim, 1987). In order to demonstrate construct validity, one must show not only that the test correlates highly with other variables with which it should theoretically correlate (convergent), but also that it does not correlate significantly with variables (discriminant) from which it should differ (Anastasi, 1988).

The Conners' Behavior Rating Scales

Development

The Conners' Rating Scales (Conners, 1969), one of the most popular checklists in use today, is the focus of this research. Collectively, two Conners' instruments are used to characterize patterns of child behavior. The Conners' Teacher Rating Scales (CTRS) and the Conners' Parent Rating Scales (CPRS). The Conners' scales are among the most widely used instruments in the schools for assisting in the evaluation and placement of children. These scales are used worldwide and have been translated into many languages.

Conners (1990) states that the original scales were developed by unknown clinicians at Johns Hopkins Hospital. The scales were used informally and qualitatively, often as extended forms of interview. Lacking a formal title, the scales were often referred to as "The Conners Scales." The Conners' Rating Scales were originally developed to provide a valid and sensitive description of a child's behavior to be used in clinical trials of various drugs and childhood hyperactivity (Conners, 1990). Today these
scales are still used mostly in the evaluation of general behavior disorders of children exhibiting possible hyperactivity or attention deficit disorders. Forms are available to both the classroom teacher and parent, making it possible to integrate several different behavioral views of the child. The two forms allow comparison between parent and teacher observations. The Conners' is quick, simple, and inexpensive to administer.

**Forms and Scales**

The Conners' Teacher Rating Scales and Conners' Parent Rating Scales both have two forms, a long form (CTRS-39 item and CPRS-93 item) and a short form (CTRS-28 item and CPRS-48). For the purpose of this study, the focus will be on the forms most recommended for use as stated in the Conners' manual (1990, p. 3): "It is suggested that the revised versions of Conners' Parent Rating Scales (CPRS-48) and Conners' Teacher Rating Scales (CTRS-28) be used for many types of clinical diagnostic applications."

Norms for the CPRS-48 were based on a sample of 529 children ages 3-17, approximately half were male. Norms for the CTRS-28 were based on a subset of the CPRS-48 sample, 383 children were drawn. Separate norms are reported by sex and age range of the child (Goyette, Conners, & Ulrich, 1978).

The CTRS-28 include the following scales: (a) Conduct Problem; (b) Hyperactivity; and (c) Inattentive-Passive. The CPRS-48 includes factors for: (a) Conduct Problem; (b) Learning Problem; (c) Psychosomatic; (d) Impulsive-Hyperactive; and (e) Anxiety. Symptoms are rated on a 4-point scale. Raw score on each factor are turned into T scores ($M = 50$, $SD = 10$). The Conners' manual notes, T scores more than or equal to 2 standard deviations above the mean could note problem areas.
Reliability and Validity of the Revised Conners' Scales

Reliability

As with any test, Sattler (1988) notes, "its properties should be thoroughly evaluated examining reliability and validity." As mentioned earlier, if one expects the results from an instrument to be useful, one needs to demonstrate acceptable reliability as well as validity. Concerning reliability, the Conners' manual states (1990, p. 38), "It should be noted that the results of this section are presented in general terms, in many cases, because individual investigators have chosen to form their own scales from the CRS...."

Inter-rater reliability of the CTRS-28 and CPRS-48 has been examined on two occasions. The first examination was of parent ratings, where product moment correlations were found between mothers' and fathers' ratings on the CPRS-48 ranging from .46 on the Psychosomatic factor to .57 on the Conduct Problem factor, the mean correlation being .51. All correlations were significant at the .001 level (Goyette, Conners, & Ulrich, 1978). The second study found the following correlations between parent and teacher ratings: Conduct Problem, $r = .33$; Learning Problem/Inattentive-Passive, $r = .45$; Impulsive-Hyperactive/Hyperactivity, $r = .36$. All correlations were significant at the .001 level. Parents consistently rated their children as exhibiting more problems ($p < .001$) with the exception of the learning problem-- Inattentive-Passive scale (Goyette, Conners, & Ulrich, 1978).

Only one study has been found concerning Internal Consistency on the CPRS-48. Total item correlations range from .13 for item 44 (vomiting or nausea) to .65 for item 6.
(sucks or chews thumb, clothing, blanket) (Goyette et al., 1978). All correlations reached the .01 level of significance (two-tailed probability). Five items (#2, 16, 24, 37, and 45) revealed statistically significant differences between mother and father ratings ($p < .01$), with mothers reporting more observed problems than fathers.

Conners (1990, p. 38) states, "No studies to date have examined the test-retest reliability of the CTRS-28, but it is expected that the results should be similar to the results obtained with the CTRS-39 ....Although factorial stability of the CPRS-48 appears adequate over time, to date there have been no published studies of its test-retest reliability." Although the Conners' manual states there have been no studies concerning test-retest reliability on either the CPRS-48 or the CTRS-28, one study was done by Diamond and Deane (1990) examining the effects and implications of frequent questionnaire administration. They found test-retest correlations greater than .87 over seven different weekly measurements using the CTRS-28. No other studies were found concerning any other type of reliability.

The revised forms of the Conners' Rating scales show few studies done when dealing with inter-rater reliability or internal consistency. However, one area of reliability appears especially weak: test-retest reliability. If this measure is used to note the temporal stability of the test over a period of time and the short form of a test has had no studies done on it, one cannot assume results would be similar to those of a longer form.

Regarding the Conners' Scales, Cohen (1988, p. 82) states, "The authors felt that clinicians choosing to adopt the abbreviated version could do so without significant loss of information....Normally, factors with few items have low reliability, resulting in limited
clinical utility." To add to doubts regarding the Conners' short forms, the original (or longer forms) have scant evidence of their own reliability.

**Validity**

At this time, face validity appears to be the strongest validity indicator when one considers the appearance, ease of administration, and widespread use (even international) of the instrument. Conners (1990, p. 39) states, "the weight of the voluminous literature is so that the scales assess important constructs which have strong explanatory and predictive utility in the characterization of childhood problem behavior." Nonetheless, predictive validity studies are not available for the CTRS-28 or the CPRS-48 at this time. Examining construct validity, no studies can be found concerning either discriminant or concurrent validity on the CTRS-28 or CPRS-48. "To say that a scale, or instrument is valid rests upon the weight of accumulated evidence from a number of validity studies using various methodologies" (Campbell & Fiske, 1959, p. 82).

**Research Summary**

There are no published studies to date examining the test-retest reliability of the Conners' Parent Rating Scale - 48. Only one study has been done on the Conners' Teacher Rating Scale-28. It was measured over consistent weekly intervals, a practice which is highly susceptible to carry-over effects. Even the Conners' studies that were noted in support of the longer forms were often done on modified versions of the Conners' or versions that were translated into different languages. As this paper has noted, for any type of assessment tool to be of use to the examiner, it must have strong reliability and validity. Despite its widespread use in the assessment of children and its obvious
popularity, the Conners’ psychometric properties appear scant on the basis of Sattler.

The stability of the Conners' scales is in great need of research.

Statement of Purpose and Hypothesis

Purpose

The purpose of this study is to examine the test-retest reliability over a two-week period of the Conners' Parent Rating scale-48 on a group of children ages 6-11.

Hypothesis

1. Using the Conners' Parent Rating Scales-48, parent ratings of their own children on each scale will significantly correlate with ratings from a re-administration two weeks later.

2. There will be no significant changes in ratings over two weeks based on the gender of the subject.

3. There will be no significant changes in ratings over two weeks based on the gender of the rater.

4. There will be no significant changes in ratings over two weeks based on the age of the subject.
CHAPTER 2

METHOD

Participants

Participants in this study were 59 elementary school children (n = 32 males, n = 27 females) ages 5-10 (M = 7.42, SD = 1.42) who were rated by a parent. The subjects were drawn from several different YMCA after-school programs in northern Texas. Raters included 42 mothers and 17 fathers (N = 59). Ninety percent of the subjects were Caucasian (n = 49), nine percent were Hispanic (n = 9) and one percent were Asian (n = 1). Fifty percent of the reported yearly family incomes involved in this study were $50,000 or above (n = 30). Twenty-nine percent of the reported yearly family incomes that participated in this study were between $35,000-$49,999 (n = 17). Fifteen percent of the families reported a total yearly income between $25,000-$34,999 (n = 9) and six percent reported incomes below $25,000 (n = 3).

Materials

The test-retest reliability of the Revised Form of the Conners' Parent Rating Scale (CPRS-48) was the instrument examined in the study. The CPRS-48 is an instrument used to characterize the behaviors of a child and compare them to levels of appropriate normative groups. The CPRS-48 is rated by the child's parent. The 48 items are rated with four responses (not at all, just a little, pretty much, very much). Responses are coded 0, 1, 2, or 3. The CPRS-48 includes scales for: a) Conduct Problem; b) Learning
Problem; c) Psychosomatic; d) Impulsive-Hyperactive; e) Anxiety; and a Hyperactivity Index. The Hyperactivity Index and the scale called Hyperactive are not synonymous. The Hyperactivity Index, composed of items most sensitive to drug effects, provides an assessment of the extent to which a child performs behaviors which are considered indicative of an underlying diagnosis of hyperkinesis.

The paper and pencil forms of the Multi-Health Systems Quik-Score Form (a quick scoring form for the Conners' tests) were used. After completion, the Quickscore Form was used to score the test and transform raw scores into T-scores. The Quickscore Forms included all of the necessary materials needed for administration, scoring, and profiling the CPRS-48.

Procedure

Upon receiving consent from the YMCA and site child-care director, a letter of information (see Appendix A) and verbal instructions were given to each participating parent along with a Conners' Parent Rating Scale-48 form. The child's parent was asked to fill out the forms on site at the time that they were given. Each form had an identification number and general information about the child and the rater. The parents were instructed to fill out the forms with a soft lead pencil or a ball-point pen. They were informed that the purpose of the rating session was to develop a complete and accurate characterization of any problems that they think they may observe in the child. The parents were encouraged to carefully remember the child in a variety of different situations before making the rating.
The parent was told to carefully read each item, consider how frequently it applies to the child, then make a rating using one of the four response options provided. The parent was told to circle each response and that if a mistake was made to simply put an "X" over the incorrect answer and circle the correct. The importance of filling out every question was stressed to the parent. The parents were encouraged to make notations in the form margins. Two weeks after the date of the first rating, at approximately the same time of day, the parents were asked to complete another CPRS-48.

Once all the data had been collected, each form was examined for unanswered responses or ambiguous ones (response sheet is blurred, etc.). The data was then be scored following the directions in the Conners' Rating Scales Manual (1990).

Data Analysis

**Hypothesis 1.** There will be high reliability coefficients for each of the CPRS-48 scales.

**Hypothesis 1 Analyses.** A Pearson product-moment correlation was computed for each of the 6 scales.

**Hypothesis 2.** There will be no significant change in ratings over two weeks based on the gender of the subject.

**Hypothesis 3.** There will be no significant change in ratings over two weeks based on the gender of the rater.

**Hypothesis 4.** There will be no significant change in ratings over two weeks based on the age of the subject.
Hypothesis 2, 3, & 4 Analysis. Since there were only two independent variables and several dependent variables a Hotelling T Square Test was used to see if the groups differed on the two dependent variables combined. The Wilks' Lambda was used to assess main effects or interactions.

Summary

This chapter has outlined the methodology used in this study. The CPRS-48 was given to the one parent of 59 elementary school children and then readministered after a period of two-weeks. Data analyses, which are reported in the next chapter, include a Pearson product-moment correlation for each of the 6 CPRS-48 scales and Hotelling T Square Tests to examine if groups differed on the two dependent variables combined.
CHAPTER 3

RESULTS

Test-retest Reliability Results

Hypothesis 1. The first hypothesis predicted that there would be a high reliability coefficient for each of the CPRS-48 scales. Pearson product-moment correlations were calculated between the first and second administrations of the CPRS-48 for each of the 6 scales. Individual scale reliability coefficients ranged from .75 to .92 (see Table 1). The reliability coefficients were all significant at the .01 level.

Assessment of Differences

Hypothesis 2. The second hypothesis predicted that there would be no significant change in the consistency of ratings over two weeks based on the gender of the subject. The sample group included 27 females and 32 males (N = 59). A Hotelling T Square Test was calculated, and the results indicated no significant differences exist between the consistency of the scale results over two-weeks and the gender of the subject $E(6,52) = .36, p = .90$ (see Table 2). A Wilks’ Lamba was calculated to assess any main effects and found no significant results ($\Lambda = 0.96$).

Hypothesis 3. The third hypothesis predicted that there would not be a significant changes in the consistency of ratings over two weeks based on the gender of the rater. The sample group included 42 females and 17 males (N = 59). A Hotelling T Square Test was calculated, and the results indicated no significant differences exist between the
consistency of the scale ratings over time and the gender of the rater $F(6,52) = 1.7, p = .15$ (see Table 3). A Wilks' Lambda was calculated to assess any main effects and found no significant results ($\Lambda = 0.84$).

Table 1

**Reliability Coefficients for the Scales of the CPRS-48**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Initial Rating</th>
<th>Second Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Conduct Problem</td>
<td>4.03</td>
<td>3.05</td>
</tr>
<tr>
<td>Learning Problem</td>
<td>2.75</td>
<td>2.45</td>
</tr>
<tr>
<td>Psychosomatic</td>
<td>0.75</td>
<td>1.14</td>
</tr>
<tr>
<td>Impulsive-Hyper.</td>
<td>4.66</td>
<td>2.93</td>
</tr>
<tr>
<td>Anxiety</td>
<td>2.56</td>
<td>2.08</td>
</tr>
<tr>
<td>Hyperactivity Index</td>
<td>7.56</td>
<td>4.85</td>
</tr>
</tbody>
</table>

**Hypothesis 4.** The fourth hypothesis predicted that there would not be a significant changes in the consistency of ratings over two weeks based on the age of the subject. The subjects were broken down into six-year-olds ($n = 19$), seven-year-olds ($n = 10$), eight-year-olds ($n = 14$), and nine-year-olds ($n = 16$). A Hotelling T Square Test was calculated and the results indicated no significant differences were found between the consistency of the scale ratings over time and the age of the subject (see Table 4). The multivariate test statistic used was the Wilks Lambda, $\Lambda = .73$, $F(18,141) = .92, p = .55$. 
Table 2

**Hotelling’s T Square Test for Sex Difference of Subjects**

<table>
<thead>
<tr>
<th>Scale</th>
<th>SS</th>
<th>MSE</th>
<th>E</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conduct</td>
<td>.09</td>
<td>1.50</td>
<td>.06</td>
<td>.81</td>
</tr>
<tr>
<td>Learning</td>
<td>.01</td>
<td>1.10</td>
<td>.01</td>
<td>.91</td>
</tr>
<tr>
<td>Psychosomatic</td>
<td>.18</td>
<td>0.43</td>
<td>.42</td>
<td>.52</td>
</tr>
<tr>
<td>Impulsive-Hyper.</td>
<td>.00</td>
<td>2.20</td>
<td>.00</td>
<td>.96</td>
</tr>
<tr>
<td>Anxiety</td>
<td>.06</td>
<td>2.03</td>
<td>.03</td>
<td>.87</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>3.27</td>
<td>4.85</td>
<td>.67</td>
<td>.41</td>
</tr>
</tbody>
</table>

Table 3

**Hotelling T Square Test for Sex Difference of Raters**

<table>
<thead>
<tr>
<th>Scale</th>
<th>SS</th>
<th>MSE</th>
<th>E</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conduct</td>
<td>4.76</td>
<td>1.42</td>
<td>3.36</td>
<td>.07</td>
</tr>
<tr>
<td>Learning</td>
<td>1.52</td>
<td>1.08</td>
<td>1.41</td>
<td>.24</td>
</tr>
<tr>
<td>Psychosomatic</td>
<td>.06</td>
<td>.44</td>
<td>.14</td>
<td>.71</td>
</tr>
<tr>
<td>Impulsive-Hyper.</td>
<td>.21</td>
<td>2.21</td>
<td>.09</td>
<td>.76</td>
</tr>
<tr>
<td>Anxiety</td>
<td>7.34</td>
<td>1.90</td>
<td>3.85</td>
<td>.06</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>1.73</td>
<td>4.89</td>
<td>.36</td>
<td>.55</td>
</tr>
</tbody>
</table>
Table 4

**Hotelling T Square Test for Age Difference**

<table>
<thead>
<tr>
<th>Scale</th>
<th>SS</th>
<th>MSE</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conduct</td>
<td>3.01</td>
<td>1.50</td>
<td>.67</td>
<td>.58</td>
</tr>
<tr>
<td>Learning</td>
<td>7.02</td>
<td>1.01</td>
<td>2.30</td>
<td>.09</td>
</tr>
<tr>
<td>Psychosomatic</td>
<td>.76</td>
<td>.44</td>
<td>.59</td>
<td>.63</td>
</tr>
<tr>
<td>Impulsive-Hyper.</td>
<td>1.71</td>
<td>2.25</td>
<td>.25</td>
<td>.86</td>
</tr>
<tr>
<td>Anxiety</td>
<td>7.90</td>
<td>1.96</td>
<td>1.34</td>
<td>.27</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>18.05</td>
<td>4.76</td>
<td>1.26</td>
<td>.30</td>
</tr>
</tbody>
</table>

**Summary**

The results of the test-retest reliability for the CPRS-48 scales produced reliability coefficients that ranged from .75 to .92. The reliability coefficients were all significant at the .01 level (see Table 1). Results from a sample group of 27 females and 32 males (N = 59) indicated no significant change in ratings over two weeks based on gender of the subject. Statistical results also indicated, with a sample of 42 females and 17 males (N = 59), no significant change in ratings over two weeks and the gender of the rater. When subjects were broken down into age groups of six-year-olds (n = 19), seven-year-olds (n = 10), eight-year-olds (n = 14), and nine-year-olds (n = 16), results indicated no significant differences were found between the consistency of the scale ratings over two weeks and the age of the subject.
CHAPTER 4

DISCUSSION AND IMPLICATIONS

Test-Retest Reliability of the CPRS-48

The purpose of this study was to examine the test-retest reliability of the Conners’ Parent Rating Scale-48 over a two-week period of time. To date there have been no published studies examining the test-retest reliability of the CPRS-48. A correlation coefficient of at least .90 is recommended for instruments used in making placement decisions, and possibly a lower one of .75 for screening procedures (Lichtenstein & Ireton, 1984). This study found test-retest reliability of the CPRS-48 scales to range from .92 to .75 over a two-week period (see Table 2). The results of this study found all of the CPRS-48 scales to meet the criteria for adequate reliability coefficients.

This study also examined the consistency of scale results over two weeks and various independent variables. It was hypothesized that no significant differences would be found when examining gender of the subject, gender of the rater, or age of the subject and consistency of the scale scores. Results from this study indicate that there are no significant differences in the consistency of how boys were rated over time compared to girls (see Table 2). There also appear to be no significant differences in how male raters rate compared to female raters, more specifically fathers and mothers (see Table 3). Lastly, this study found no significance in the age of the child and the consistency of their rating over time (see Table 4).
Result Implications

The CPRS-48 is a behavior rating scale that is one of the most widely used instruments in the schools for assisting in the evaluation and placement of children. A behavior checklist is intended to reflect a complex interaction among the following: the scale or checklist used (scale values, wording of items, content of items, standardization sample); the child (age, gender, type of disturbance); the informant (expectancies, recall ability, response bias, comprehension of items); the examiner (gender, ethnicity, traits); the setting (school, home, playground, hospital, prison); and the reasons for the evaluation (screening, diagnosis, placement, program evaluation) (Sattler, 1988). In order to accurately reflect the above-mentioned interactions, a measure must first be proven consistently valid and reliable over time.

The present study produced a small piece of evidence concerning the consistency of the CPRS-48. However, the numbers used in each cell group for the comparison studies were very small, making the power of the test results relatively weak. Aside from a small total sample number, most of the families used in this study were Caucasians living in middle-class communities. In comparison to the evidence that must be accumulated in order to call the Conners' Parent Rating Scales valid and reliable, this one examination of its reliability over a two-week period of time must be considered only the beginning of much-needed empirical support. More evidence must be accumulated before these scales may be considered valid and reliable. At this time, the Conners' Scales should be used as a subject of further research not considered a valid form of child assessment.
Further Research and Development

In order to thoroughly evaluate the integrity of a checklist, one must review all aspects of an instrument's reliability and validity. The present research provides only a small piece of what is needed to create an existing network of supportive psychometric evidence on the CPRS-48. When evaluating any checklist one must investigate: (a) the characteristics of its standardization sample, (b) the types and quality of its reliability, (c) the reliability and validity of comparable tests, (d) the scoring procedures, (e) the method of administration, (f) the limitations, and (g) the strengths (Kaplan & Saccuzzo, 1982).

Despite its widespread use in the assessment of children, the CPRS-48 has very little research to support its psychometric properties. The Conners' Scales, in general, have very poor evidence of research-supported integrity. Future studies are needed in order to produce a strong foundation of support for the Conners' Scales. Conners (1990, p. 39) himself recognized the importance of instrument integrity when he noted, "To say that a scale, or instrument is valid rests upon the weight of accumulated evidence from a number of validity studies using various methodologies."

Summary

The current study produced adequate test-retest correlation coefficients for the CPRS-48 scales ranging from .92 to .75. This study found no significant changes in scores over two weeks regardless of the gender of the subject, gender of the rater, or age of the child. The results of this research lend support to the integrity of the test-retest reliability of the CPRS-48. Very little research has been done on the psychometric properties of the CPRS-48 or any of the three other Conners' Rating Scales. Despite the widespread use
and popularity of the Conners’ Scales, substantially more evidence is needed concerning instrument reliability and validity.
APPENDIX A

PARENT FORM
Dear Parent:

I'm a graduate student conducting a study designed to examine the Conners' Behavior Rating Scales. I'm asking you to help me with this study by filling out the following forms. This study will help us learn what tests provide the best behavioral information on children. I'm hoping that the results of this research will benefit teachers and parents like yourself in providing a smaller amount of questions to answer, but the same amount of valuable information on your child as the longer forms provide. Valid testing is essential in helping your child.

This study consists of two identical test forms that I will ask you and your child's teacher to fill out on two separate occasions (about two weeks apart). You will also be asked to reply to the questions below by checking the area that best describes your answer. The information that you will be offering by participating will help in the collection of very valuable information concerning the strengths of the Conners' tests.

Your child's results will be kept confidential. At the conclusion of this study, a summary of group results will be made to all interested parents and teachers. This study has been approved by the University of North Texas Committee for the Protection of Human Subjects. Thank you for your cooperation and support.

Sincerely,

Debra Gomez
University of North Texas

THIS PROJECT HAS BEEN REVIEWED BY THE UNIVERSITY OF NORTH TEXAS COMMITTEE FOR THE PROTECTION OF HUMAN SUBJECTS.
Please check the answer that best applies:

1. Your relationship to the child.
   Mother __ Father __ Grandmother __ Grandfather __
   Other (please specify) ______________________

2. The child's sex (circle one): Boy   Girl

3. The child's age ______

4. The child's race (circle one): Caucasian Hispanic
   African American Native American Asian
   Other: __________

5. Your race (if different from the child): __________

6. Which group best describes your family income per year?
   under $10,999 __ $11,000-$15,000 __ $16,000-$24,999 __
   $25,000-$34,999 __ $35,000-49,999 __ Over 50,000 __
REFERENCES


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


