1997

The Stability of WISC-III Scores: For Whom are Triennial Re-Evaluations Necessary?

Linda C. Cassidy
University of Rhode Island

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation
https://digitalcommons.uri.edu/oa_diss/933

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons@etal.uri.edu.
THE STABILITY OF WISC-III SCORES:

FOR WHOM ARE TRIENNIAL RE-EVALUATIONS NECESSARY?

BY

LINDA C. CASSIDY

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

1997
ABSTRACT

This study investigated the long-term stability of Verbal, Performance and Full Scale IQ scores for a sample of exceptional children evaluated with the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) as part of the mandated three-year re-evaluation process. Archival data were collected from the special education files of 592 children who were administered the WISC-III on two separate occasions between September, 1992 and June, 1996 from twenty-seven school districts in a small New England State. Several variables (special education classification, Full Scale IQ score at initial administration, age of participant at initial administration, as well as the ethnicity and gender of the participants) were examined to detect which variables, if any, would influence the stability of test scores over the three-year interval. Three statistical analyses for measuring stability (correlational method, test of mean difference and test of intra-individual variability) were conducted. Results indicated that for this sample of children in special education, examined as a group, Verbal IQ, Performance IQ and Full Scale IQ scores remain stable over time. However, certain populations demonstrated significant instability in scores over the three-year interval. Results suggested that children classified as mentally retarded or behavior disordered fluctuate significantly in their performance on the WISC-III over a three-year interval. Similarly, children who receive an initial Full Scale IQ score above 109 demonstrate significant instability in scores between administrations. Therefore, for these populations, three-year re-evaluations appear necessary. However, routine administrations of the WISC-III for all children involved in special education is of questionable value. Further research must be done to confirm the
findings of this study in order to assist policy makers in distinguishing which children
would benefit from a three-year re-administration of the WISC-III and for which children
such information would yield no further information.
ACKNOWLEDGMENT

It is not often that a student is fortunate enough to have such a supportive dissertation committee as I had. I would like to thank each committee member for all their time and encouragement that they have given me. To Dr. Janet Kulberg, my major professor: it was very difficult and challenging to complete my dissertation during my internship year, especially out-of-state. However, you guided me throughout the entire process. I sincerely thank you for your advice, time and commitment not only to my research but also to my professional development. To Dr. Paul Florin and Dr. David Byrd: thank you for all your support and encouragement while completing this project. I would also like to thank Dr. Denise De Zolt. You supported, advised and motivated me throughout my graduate career at the University of Rhode Island. Your assistance in the acquisition of funding for this project is greatly appreciated and your dedication to the profession and to your students is admired.

I would also like to thank my family for all their love and support. Each of you have given me the encouragement and the motivation to succeed.

Without the support of Jim, the most important person in my life, I would have left graduate school on the first week. Because of your unselfish nature, my goals have been reached. The degree I have received is as much yours as it is mine.
# TABLE OF CONTENTS

## I. Introduction
- Rationale for Study .............................................................. 1
- Review of the Literature ...................................................... 2
- Statement of Hypotheses ...................................................... 37

## II. Method
- Participants .................................................................................. 40
- Instruments ................................................................................... 41
- Procedure ....................................................................................... 41

## III. Results
- Stability of IQ Scores for the Entire Sample ................................ 43
- Stability of IQ Scores Within Specified Populations .................... 45
- Special Education Classification .................................................. 46
- Initial Full Scale IQ Score ............................................................ 51
- Age at Initial Administration ....................................................... 54
- Ethnicity of Participants ............................................................. 57
- Gender of Participants .................................................................. 59

## IV. Discussion
- Stability of IQ Scores for the Entire Sample ................................ 62
- Stability of IQ Scores Within Specified Populations .................... 63
- Special Education Classification .................................................. 64
- Initial Full Scale IQ Score ............................................................ 69
- Age at Initial Administration ....................................................... 71
- Ethnicity of Participants ............................................................. 73
- Gender of Participants .................................................................. 74
- Conclusions ................................................................................... 75
- Limitations ..................................................................................... 76
- Implications ................................................................................... 77

## V. Appendices
- Appendix A- Institutional Review Board Letter of Approval ......... 80
- Appendix B- Correspondence ....................................................... 82

## VI. Bibliography ............................................................................ 89
LIST OF TABLES

1. Means and Standard Deviations for Verbal IQ, Performance IQ, and Full Scale IQ Scores for the Entire Sample................................................................. 44

2. Means, Standard Deviations, Correlation Coefficients, and Tests of Mean Differences (t) Between Time 1 and Time 2 Based on Initial Special Education Classification................................................................. 47

3. Means, Standard Deviations, Correlation Coefficients, and Tests of Mean Differences (t) Between Time 1 and Time 2 Based on Full Scale IQ Score at Initial Administration................................................................. 52

4. Means, Standard Deviations and Correlation Coefficients For Time 1 and Time 2 Based on Age of Participants at Initial Administration........................................................................................................ 55

5. Means, Standard Deviations and Correlation Coefficients For Time 1 and Time 2 Based on Ethnicity of Participants................................................................. 58

6. Means, Standard Deviations and Correlation Coefficients For Time 1 and Time 2 Based on Gender of Participants................................................................. 60
Chapter 1
INTRODUCTION
Rationale for Study

Criteria for determining eligibility for special education services under federal and state legislation requires the administration of an intelligence test to measure a child's cognitive functioning. This has generally been interpreted as an individually administered, standardized intelligence test. Such a measure is re-administered every three years to determine if the child is still eligible for special education services. Therefore, results of the individual intelligence test is a viable component in the decision making of continuation of services.

The Wechsler Intelligence Scale for Children- Third Edition (WISC-III) is the current cognitive test of choice for most practicing school psychologists. There have been no studies at this point, however, that address the stability of the WISC-III IQ scores over the mandated three-year period. Therefore, there is a necessity to investigate the stability of the WISC-III in order to examine the reliability of the scores that are used frequently in the decision making process regarding a child's classification and eligibility of services received. Such an investigation will assist in determining whether there are benefits in re-administering the same test to the same child during the three-year re-evaluation process.

If one accepts the assumption that IQ scores remain relatively stable over time, then re-evaluations appear unjustified. However, if it is determined that instability of scores is expected for certain populations, then it is imperative to determine for which children scores will remain stable over time and for which children score fluctuations will be expected.
The purpose of this study is to investigate the long-term reliability and stability of the Verbal, Performance and Full Scale IQ scores for a sample of exceptional children. In addition, several variables (special education classification, Full Scale IQ score at initial administration, age of the participants at initial administration, ethnicity and gender) will be examined to detect which variables may influence the stability of test scores over the mandated three-year re-evaluation interval.

Research addressing this area is lacking and the importance of conducting this research will greatly influence the decision making practices of school personnel in determining classification and placement of children as a result of a three-year re-evaluation.

Review of the Literature

The Nature of Intelligence

In order to understand the role of intelligence testing in the special education community, it is important to first define the nature of intelligence. Currently, there is no single comprehensive definition of the construct, intelligence, that is accepted throughout the scientific community (Sattler, 1988). This is partially due to the development of various definitions of intelligence derived from several theoretical orientations.

An early and influential theory developed to define the nature of intelligence was proposed by Charles Spearman. In his 1927 publication, The Abilities of Man, Spearman defined his two-factor theory of intelligence. According to Spearman, intelligence is comprised of two principal components. One component of intelligence is designated as the general factor and denoted by the letter g...

it is so named because, although varying freely from individual to individual, it remains the same for any one individual in respect of all
the correlated abilities. The second factor has been called the 'specific factor' and denoted by the letter $s$. It not only varies from individual to individual, but even for any one individual from each ability to another (p. 75).

According to Spearman's theory, the $g$ factor is related to perceiving and manipulating relationships, thus representing abstract reasoning. Because performance on various tasks did not correlate perfectly with $g$, Spearman suggested that each task includes a specific ability or $s$ ability. Spearman's theory greatly influenced the development of David Wechsler's conceptualization of intelligence and the development of the Wechsler scales. Wechsler ascribed to the phenomenon of $g$ and defined intelligence as the "overall capacity of the individual to understand and cope with the world around him" (Wechsler, 1974). Although Wechsler's scales have been dichotomized into verbal and performance scales, Wechsler adhered to a strong emphasis on the assessment of $g$ or general intelligence. In this manner, the Performance and Verbal scales are "primarily a way of identifying two principal modes by which human abilities express themselves" (Wechsler, 1974).

Intelligence, according to Wechsler (1974), is a multidimensional, global entity rather than an independently defined psychological trait. Thus, intelligent behavior is the combination of qualitatively different abilities for the purpose of thinking rationally, logically and dealing effectively with the external world.

A second influential theory of intelligence was proposed by L.L. Thurstone (1938). Thurstone suggested that intelligence consists of independent abilities, not a single general factor, as proposed by Spearman. Some of these primary mental abilities consist of verbal comprehension, word fluency, perceptual speed and reasoning. According to Thurstone's
theory, individual differences exist among abilities, whereby individuals demonstrate more proficiency in some areas and not in others.

A recent theory of intelligence proposed by Robert Sternberg (1985) stresses the importance of the manner in which individuals solve problems instead of the answers that are derived. This triarchic theory of intelligence emphasizes the examination of the mental activities individuals engage in while solving problems.

Although there are several theories that attempt to explain the nature of intelligence, this construct has currently been synonymous with results received from intelligence measures. Therefore the development of the intelligence test has a great importance in the understanding of the nature of intelligence.

The History of Intelligence Testing

Measuring intelligence is an abstract procedure. Currently, the most effective way to measure intellectual abilities is indirectly, by evaluating and observing the intellectual behaviors of an individual as demonstrated through tasks on an intelligence test.

The development of intelligence testing has been made possible by the contributions of early experimental psychologists (Anastasi, 1988). Early work by Wilhelm Wundt in 1879 resulted in the discovery of individual differences in sensory abilities and reaction time. These measures were later instituted as part of formal cognitive measures.

Sir Francis Galton, the father of mental tests, developed the first assessment designed to measure intelligence in the late nineteenth century (Boring, 1950). Galton believed that one’s understanding of the environment is a result of how the information is received through the senses. Therefore, individuals with high intelligence would have exceptional sensory discrimination abilities. As a result, Galton developed a battery of tests to
examine sensory discrimination and motor coordination in order to assess mental functioning. However, none of the traits measured correlated with his theory of intelligence.

Similar work was conducted in the United States during this period of time, namely through the work of James McKeen Cattell. Like Galton, Cattell believed that simple sensory, perceptual and motor responses were the key dimensions of intelligence (Boring, 1950). Cattell and Galton provided little conclusions about intelligence. However, as a result of their investigations many questions were raised about the nature of intelligence that are currently being examined today.

The first standardized intelligence test was devised by two French psychologists, Alfred Binet and Theophile Simon (Doll, 1962). In 1904, the French Ministry of Education commissioned Binet to develop a formal assessment procedure to measure the cognitive performance of children. The underlying purpose was to determine which children would be placed in special programs for the ‘mentally deficient.’ Binet and Simon produced an objective test to measure the skills necessary for academic success such as mathematical skills, memory, language abilities and the ability to follow directions. It was believed that by using such an array of items, the test would inevitably tap enough abilities to assess a child’s intellectual potential. The contribution of Binet and Simon’s work has had a lasting effect on special education eligibility. Although test development and standardization procedures have improved over time, “Binet’s early work has had considerable influence on adoption of the ‘diagnostic/prescriptive’ paradigm that has guided special education for much of its history in the United States” (Meyen, 1995).
Lewis Terman, at Stanford University, revised Binet's assessment procedure and standardized it with American children in 1916 as the Stanford-Binet Intelligence Scale which met high psychometric properties for the time (Terman, 1916). In addition, the revised Stanford-Binet met Terman's call for intelligence tests that not only identified a global estimate of cognitive ability (g) but also measured specific abilities (s) as well.

David Wechsler published his first intelligence test for children in 1949 which was referred to as the Wechsler Intelligence Scale for Children (WISC). Two revisions have since been published, the Wechsler Intelligence Scale for Children- Revised (WISC-R) and the Wechsler Intelligence Scale for Children- Third Edition (WISC-III). The WISC-III was recently published in 1991 and is currently the test of choice for most practicing school psychologists.

The Practice of Re-evaluations

Measuring the intelligence of a child with special needs is a necessary component in determining the eligibility for special education services. According to a survey conducted by the National Association of School Psychologists (NASP) in 1987, practicing school psychologists spend two-thirds of their time engaged in issues related to special education eligibility determination (Reschley, Genshaft & Binder, 1987). The utilization of intelligence testing for school psychologists is so prevalent that according to this NASP survey, school psychologists administer close to eighteen intelligence measures per month. The purpose of the administration of measures of intelligence has generally been interpreted to assist not only in eligibility determination, but also play a key role in guiding appropriate placement and assessing the individual academic needs of the child (Galvin & Elliott, 1985).
According to PL 94-142, schools are not only required to assess eligibility for special education services, but are also mandated to re-evaluate children involved in special education every three years. The implicit rationale of this mandate is to confirm the original diagnosis and placement as well as to reassess the academic needs of the child (Galvin & Elliott, 1985).

The re-evaluation process typically includes a repetition of an initial standardized test battery (Elliott, Piersel & Galvin, 1983; Hartshorne & Hoyt, 1985). Thus, for many school psychologists the re-evaluation is quite indistinguishable from an initial evaluation in both the purpose of the assessment as well as the procedure. Although three-year re-evaluations are mandated at this time, there are no requirements under PL 94-142 nor State guidelines that a routine battery of standardized tests must be administered (Safer & Hobbs, 1980; Hartshorne & Hoyt, 1985). Therefore, many school psychologists question the need for the re-administration of an entire battery of individualized standardized assessments. Because federal legislation does not require the re-administration of an individualized cognitive assessment, it is essential for school psychologists to determine the appropriate procedures and tools that would yield the most useful information in determining the needs of the child concerning placement and services received.

According to Anderson, Cronin and Kazmierski (1989), “The re-evaluation information that is presently available has failed to provide direction for diagnosticians and special education administrators who are charged with the responsibility of meeting students' assessment needs and complying with federal evaluation requirements” (p. 941). In this manner, one must question the necessity of re-evaluations, the effectiveness of re-evaluations and the efficacy of the process.
Elliott, Piersel and Galvin (1983) conducted a study whereby results suggest that school psychologists spend 15% of their time each year performing re-evaluations, and 80% of school psychologists spend more than three hours per evaluation. Therefore, the time involved in conducting mandated re-evaluations are quite extensive. In terms of financial strain, testing costs consume almost one-third of the amount spent in educating a child in special education for one academic year (Smith, 1982).

As one questions the necessity of psychological re-evaluations, it is essential to be cognizant of the research that has addressed this issue. According to a survey conducted by Galvin and Elliott (1985), there generally is a relatively low incidence of perceived and actual change in classification and placement for children in special education as a result of a three year re-evaluation. As a result, the need for an automatic re-administration of an intelligence assessment for every child receiving special education services is questionable. While the purpose of re-evaluations appears quite logical, studies have currently raised the question of automatic three year re-evaluations in terms of the efficiency, cost effectiveness and overall value of re-testing, especially the re-administration of an individual intelligence measure.

Reliability of Intelligence Test Scores

Test reliability refers to "the consistency of scores obtained by the same persons when reexamined with the same test on different occasions" (Anastasi, 1988). Psychometric theory adheres to the belief that an individualized obtained score is composed of a true score and an error score (Sattler, 1988). The true score estimates the amount of the trait or characteristic the child actually possesses, while the error score refers to the extent to which individual differences are due to chance factors. Therefore,
measures of test reliability estimate what proportion of the variance is due to true differences and what proportion is due to error factors.

Reliability can be measured in a variety of ways. The most common computation that determines the degree of consistency in test scores is the reliability coefficient. The reliability coefficient represents a ratio of the true score variance to the observed score variance (Sattler, 1988). Reliability information is easily attained through a test-retest procedure. This procedure requires the administration of the same test to the same group of participants on two occasions and correlates the resultant test scores. This method has been shown to provide the most reasonable estimate of test reliability (Bauman, 1980). The correlation coefficient that is most often used in test-retest reliability studies is the Pearson product-moment correlation coefficient. This correlation coefficient represents the degree to which the obtained score is consistent or stable over time. Other methods that have been implemented in test-retest reliability studies are the examination of group mean differences and the examination of the existence of intra-individual variability.

Test-retest reliability represents the extent to which one's scores on a given test can be generalized over different times. Thus, the higher the reliability coefficient, the less susceptible one's scores are due to random changes in the condition of the testing environment or the condition of the test taker (Anastasi, 1988). "For most tests of cognitive and special abilities, a reliability coefficient of .80 or higher is generally considered to be acceptable" (Sattler, 1988).

The investigation of test-retest reliability of intelligence measures is essential because the implicit assumption underlying the procedures of special education eligibility and placement is heavily based upon their anticipated stability over time (Webster, 1988). For
this reason, exploring the test-retest reliability of the WISC-III, which is the current test of choice, would be beneficial in understanding the meaningfulness of scores when compared to scores achieved three years earlier.

The original Wechsler Intelligence Scale for Children (WISC) was published in 1949. A revised edition, the WISC-R, was published in 1974 and the current edition, the WISC-III, was published in 1991. Since the original WISC publication, studies have investigated (a) the stability of WISC scores, (b) the comparability of WISC and WISC-R scores, (c) the stability of WISC-R scores and (d) the comparability of WISC-R and WISC-III scores over time. No studies at this time have investigated the stability of WISC-III scores over the mandated three-year interval. Before investigating WISC-III stability, however, it is important to review the literature in the examination of the procedures previously used to investigate test-retest reliability, as well as the factors that have influenced the stability of intelligence scores over time.

Stability of the Wechsler Intelligence Scale for Children (WISC)

The original WISC was published in 1949 as a means to assess the cognitive functioning of children aged six years, zero months through sixteen years, eleven months. Since its publication, numerous studies have been conducted with "normal" children as well as with exceptional children to investigate the reliability and stability of the Verbal, Performance and Full Scale IQ scores over time. Several variables have been considered in this investigation to determine whether score stability is a factor in certain populations. Special education classification

Non-special education population. Before examining the stability of scores of a special education population, it is of importance to first investigate the stability of scores
with samples of non-special education children. A study conducted by Gehman and Matyas (1956), examined the stability of WISC scores for a group of sixty children not receiving special education services over a four year period. Test-retest reliability coefficients were .77, .74 and .77 for Verbal IQ (VIQ), Performance IQ (PIQ) and Full Scale IQ (FSIQ) scores, respectively. Conklin and Dockrell (1967) also investigated the long term stability of WISC scores over a two and four year retest interval with 101 ten year old children. Results of the two-year interval yielded coefficients of .83, .61 and .80 for VIQ, PIQ and FSIQ. Correlation coefficients of the four year interval were relatively lower (.64, .54, and .72 for VIQ, PIQ and FSIQ, respectively).

Several studies have investigated the stability of WISC scores with specific special education populations, namely children diagnosed as mentally retarded, learning disabled and emotionally disturbed.

Mentally retarded population. Throne, Schuman and Kaspar (1962) detected high test-retest reliability and stability over a three to four month interval for a group of thirty-nine mentally retarded boys aged eleven to fourteen years, eleven months. Results produced correlation coefficients of .92, .89 and .95 for VIQ, PIQ and FSIQ. In addition, no significant differences between the mean IQ scores of the two administrations were reported. After a seventeen month interval, Friedman (1970) examined the stability of WISC scores utilizing a group of forty-four children enrolled in self-contained classes for children classified as mentally retarded. Results suggested that no statistically significant differences were obtained between mean IQ scores. However, the reported correlation coefficients (VIQ=.48, PIQ=.78, FSIQ=.68) were relatively lower than the findings of Throne et al., (1962). With a longer test-retest interval, Walker and Gross (1970)
examined the stability of WISC scores with forty-nine mentally retarded children with a retest interval ranging from two to three and one half years. Results indicated reliability coefficients of .70, .73 and .76 for VIQ, PIQ and FSIQ, respectively. Rosen, Stallings, Floor and Nowakiwska (1968) used the same retest interval for forty-nine mentally retarded children and correlations (VIQ=.70, PIQ=.72 and FSIQ=.81) were similar to the results of Walker and Gross (1970). As the results suggest from this review of the literature, WISC scores remain relatively stable over time for a mentally retarded population.

Learning disabled population. Coleman (1963) examined the stability of WISC scores with a group of twenty-four boys classified as learning disabled. After a mean test-retest interval of fifteen months, correlation coefficients were .62 for VIQ, .81 for PIQ and .77 for FSIQ. In addition, significant mean differences existed between test-retest scores for VIQ and PIQ. Mean VIQ decreased two points over time and mean PIQ increased three points over time.

Emotionally disturbed population. Tigay and Kempler (1971) investigated the stability of WISC scores with twenty-five hospitalized children experiencing emotional disturbance. After a mean retest interval of 7.8 months, results produced correlation coefficients of .82, .51 and .83 for VIQ, PIQ and FSIQ respectively. In terms of mean differences, PIQ and FSIQ were significantly higher on the second administration whereas the VIQ was constant between administrations.

Examining the effects of special education classification on the stability of WISC scores, results of these studies conclude that VIQ, PIQ and FSIQ scores remain relatively
stable for children classified as mentally retarded. However, for children classified as learning disabled and emotionally disturbed, score instability was prevalent.

**Full Scale IQ at Initial Administration**

Whatley and Plant (1957) examined the stability of WISC IQ scores for a group of seventy children referred for testing who received an initial FSIQ score below 90. After a mean test-retest interval of seventeen months, no significant mean differences were found between IQ scores. Thus, the researchers concluded that FSIQ, VIQ and PIQ scores are relatively stable over time for children with initial FSIQ scores of 90 or below.

**Age at Initial Administration**

Irwin (1966) investigated the reliability of WISC scores at two selected age levels (Group I= five years, seven months to six years, six months and Group II= ten years, seven months to eleven years, seven months) over a three to five week test-retest interval. Results indicate that younger children are more variable in their ability over time than older children. Similarly, Klonoff (1972) reported in his sample of 173 children ranging in ages from five to thirteen that the youngest group of children (aged five) exhibited the most instability over a year period. Therefore, it appears that younger children experience greater instability in WISC scores over time than older children.

**Gender**

In terms of the relationship between gender and test stability, Klonoff (1972) reports no gender differences in the patterning of IQ change.

**Summary of WISC Stability Studies**

Review of the literature investigating the test-retest stability of WISC scores suggest that VIQ, PIQ and FSIQ scores remain moderately stable over time. Specifically, results
conclude that non-special education children, children classified as mentally retarded, and children who received an initial FSIQ score below 90 have stable WISC IQ scores over a test-retest interval. However, results also indicated that for certain populations, WISC IQ scores demonstrate significant instability over time. Children classified as learning disabled or emotionally disturbed significantly fluctuate in scores. In addition, younger age groups systematically elicit more instability in scores over time than older children. Therefore, it appears that certain variables (such as special education and age) may attribute to test-retest score instability of the WISC.

Comparison of WISC/WISC-R Scores

The Wechsler Intelligence Scale for Children (WISC-R) was published in 1974. The revised manual stated that "the revision of the WISC represents a synthesis of two somewhat opposing aims: (a) the retention of as much of the 1949 WISC as possible because of its widespread use and acceptance, and (b) the modification or elimination of items felt by some test users to be ambiguous, obsolete, or differentially unfair to particular groups of children" (Wechsler, 1974, p. 10). Such elimination and alteration of items were due to reports in the literature that certain items were unfair or culturally weighted (Wechsler, 1974). Other modifications were made in the new revision of the Wechsler Scale. Several subtests were lengthened to enhance reliability. The administration of the Verbal and Performance subtests were modified so that the subtests alternated. Another important change of the WISC-R was the use of updated norms and the inclusion of nonwhite children in the normative sample (Reschley & Davis, 1977).

The manual for the revised Wechsler Intelligence Scale for Children (WISC-R), did not include any empirical data regarding the comparability of scores from the WISC-R.
with scores from the original WISC (Wechsler, 1974). With no information provided in the manual, researchers were forced to conduct studies comparing the stability of WISC and WISC-R scores over time. Since the WISC-R had become widely utilized, interpretation of change scores had been the sole responsibility of practicing psychologists. Information comparing previously administered WISC scores and currently administered WISC-R scores was very important in the re-evaluation process of children receiving special education services.

Because performance on intelligence measures is a major criterion in determining eligibility for special education services, sudden changes in scores could potentially and seriously affect the special education placement of children in need of services. For example, a decrease in scores from the WISC to WISC-R will ultimately decrease the discrepancy between ability and achievement, thus decreasing the number of children eligible as learning disabled and increasing the number of children classified as mentally retarded.

There have been a number of studies that have examined the difference between WISC and WISC-R scores in a variety of populations over variable test-retest intervals. However, because research has examined the comparability of two distinct instruments, a counterbalanced design has been typically utilized to control for error, such as practice effects. Studies employing a counterbalanced design are more methodically sound. In these studies, children are randomly assigned into two groups. Half of the children are administered one edition of the WISC followed by the alternative edition while the other half is administered the alternative edition followed by its counterpart. Many studies examining the comparability of the two WISC scales have indicated a differential practice
effect. Such a practice effect results in a significant difference in IQ scores when the WISC is administered subsequent to the WISC-R (Davis, 1977; Hamm, Wheeler, McCallum, Herrin, Hunter, & Catoe, 1976; Swerdlik, 1978; Tuma, Appelbaum, & Boe, 1978).

Questions have been raised regarding the reasons for this significant difference in scores based upon the order of administration. Flynn (1984) suggests that administration procedures are quite dissimilar between editions, especially the use of direct teaching: (a) On the WISC-R the examiner explains the correct answer to the child if he/she fails the first item of a subtest; (b) The WISC-R supports an administration whereby the examiner can probe a child’s answer; (c) On questions whereby a correct solution requires two answers, the examiner is required to specifically ask the child for a second response if only one is given. These practices are not allowed in the administration of the WISC. It appears that when the WISC is administered first to a child, a practice effect operates when administered the WISC-R because of the similarity in tasks. This practice would basically cancel out due to the operation of different test norms. However, when the WISC-R is administered first to the same child, the child’s score benefits from not only the practice effects of the similar tasks but the child also benefits from direct teaching. Such benefits as practice effects and direct teaching may lead to an artificially inflated discrepancy score. Therefore in order to determine the comparability of the WISC and WISC-R, it is essential to utilize a counterbalanced procedure to control for such discrepancy due to this order effect.

Reviews of the literature investigating the reliability between WISC and WISC-R scores have been conducted by employing a meta-analysis procedure. Kaufman (1979)
reviewed the results of eight studies which employed a counterbalanced design. The sample varied according to age, diagnosis, sample size, and test-retest interval. Overall, the results indicate that the WISC-R VIQ tends to be 4.7 points lower, the PIQ 6.4 points lower and the FSIQ 5.8 points lower than the corresponding WISC IQ scores. Flynn (1984) reviewed twenty-nine studies comparing the IQ scores of the WISC and WISC-R resulting in analyses of 1607 subjects. Results concluded that the mean WISC-R Full Scale IQ score was typically 8.54 points lower than the WISC Full Scale IQ score.

**Special Education Classification**

Non-special education population. Before examining whether there are significant differences in IQ scores between the WISC and WISC-R, as attributed to special education classification, it is important to examine the studies conducted with the normal school-age population. Schwarting (1976) examined fifty-eight randomly selected children ranging in age from six to fifteen years of age at the time of the two administrations. The order of administrations was counterbalanced and the test-retest interval varied from sixty to sixty-seven days. Results indicated that the WISC-R Verbal, Performance and Full Scales IQ scores were all significantly lower than the WISC scores by 4.86, 8.74, and 7.49 points, respectively. With a similar population, Stokes, Brent, Huddelson, Rozier, and Marrero (1978) questioned the trend of lower scores on the WISC-R. Fifty-nine sixth graders were administered the WISC and WISC-R in counterbalanced order over a mean interval of forty-two days. Results also suggested that WISC-R Verbal IQ and Full Scale IQ scores were significantly lower than the comparable WISC scores. However, no significant differences were reported between Performance IQ scores. Similarly, correlation coefficients were relatively high between administrations...
(VIQ = .92, PIQ = .76, FSIQ = .89). With a slightly smaller sample of thirty-six normal Caucasian ten-year old children, Tuma, Appelbaum and Bee (1978) examined the stability of WISC and WISC-R scores. In a counterbalanced design and a mean interval of twenty days, results also indicated significantly lower WISC-R Full Scale IQ scores. Thus, it appears that a trend exists whereby normal school-aged children will score significantly lower on the WISC-R than on the original edition of the WISC.

Several studies have examined the comparability of the WISC and WISC-R by way of special education classification.

Mentally retarded population. With short test-retest intervals ranging from three days to thirty-nine days, studies reported WISC-R FSIQ scores to be significantly lower than WISC FSIQ scores for mentally retarded children. For example, Solly (1977) administered the WISC and WISC-R to twelve mentally retarded children in counterbalanced order over a three day interval. Results indicated that children scored on the average 10.8 points lower on the WISC-R FSIQ. Catron and Catron (1977) examined the WISC and WISC-R scores of sixty-two mentally retarded children each administered the WISC and WISC-R in counterbalanced order over an interval of three weeks. All three WISC-R IQ scores were significantly lower than their WISC counterparts (5 points between VIQ scores, 6 points between PIQ scores and 5.5 points between FSIQ scores). Berry and Sherrets (1975) examined the comparability of scores of twenty-eight mentally retarded children over a fourteen day period utilizing a counterbalanced design. Mean WISC-R scores were also lower than WISC scores across all three IQ measures (VIQ = 4.43 points, PIQ = 3.25 points, and FSIQ = 3.43 points). Hamm, Wheeler, McCallum, Herrin, Hunter, and Catoe (1976) examined scores of forty-eight students enrolled in a
self-contained mentally retarded class. The average test-retest interval was thirty-nine
days. Results indicated that WISC-R IQ scores were significantly lower than WISC
Verbal IQ (6 points), Performance IQ (10 points) and Full Scale IQ (7.5 points) scores.
Reschley and Davis (1977) examined the comparability of WISC and WISC-R scores
among forty-eight mentally retarded children after a mean interval of seventeen months.
Results indicated similar correlations as the 1976 study conducted by Hamm et al. (VIQ= .83, PIQ= .80, and FSIQ= .87). Results were also similar in that mean significant
decreases were established between the WISC and WISC-R for VIQ (7 points) and FSIQ
(4 points) scores. However, no significant difference was reported for PIQ scores. With a
longer test-retest interval averaging 2.3 years, Spitz (1983) examined the comparability of
the WISC and WISC-R with 282 mentally retarded children. Results indicated a FSIQ
decrease of 5.12 points from the WISC to WISC-R. The FSIQ test-retest correlation
reached .70.

Results of these studies for the mentally retarded population are in agreement with
Hamm et al.'s (1976, p. 140) observation that "...the author was aware of the earlier
maturation, greater test sophistication, and the increasing availability of manipulative
materials similar to the subtest tasks present at the WISC-R standardization in 1974.
Therefore, today's subjects who have this maturity and experience should tend, in general,
to obtain higher scores on the WISC than on the WISC-R when exposed to both scales."

As a result, it has been concluded by Hamm et al. (1976) that when the WISC-R is
used as a main criterion for classification and placement of mentally retarded children,
there will be an increase in the number of students classified as mentally retarded because
of the significant decrease in IQ scores from the WISC to WISC-R.
Learning disabled population. Paal, Hesterly & Wepfer (1979) administered the WISC and WISC-R in counterbalanced order to a group of forty students classified as learning disabled ranging in age from six to ten. After a test-retest interval of sixty to sixty-seven days, reports indicate that the WISC-R VIQ and FSIQ were both significantly lower than the WISC VIQ and FSIQ. There were no significant differences between the PIQ. In a study comprised of a larger sample of 186 learning disabled children and a mean test-retest interval of two years, results indicated no significant differences between FSIQ scores (Covin, 1977). Reliability coefficients for FSIQ ranged from .85 at age nine, to .96 at ages eight, fourteen, and fifteen. Covin (1976) also examined the comparability of the WISC and WISC-R with a group of 101 elementary school aged children with learning difficulties administered two years apart. Correlation coefficient for the FSIQ between the WISC and WISC-R was reported to be .95. However, the mean FSIQ score for the WISC-R was significantly lower than the WISC mean FSIQ score.

Several special education populations. Zimmerman and Woo-Sam (1972) examined the comparability of WISC and WISC-R IQ scores utilizing a sample of eighty-six children diagnosed as either mentally retarded or emotionally disturbed. Results indicated that for the emotionally disturbed sample, WISC scores were significantly higher in VIQ (4.9 points), PIQ (3 points) and FSIQ (4.1 points) scores as compared to WISC-R scores. For the mentally retarded sample, WISC scores were also significantly higher in VIQ (3.3 points), PIQ (2.2 points) and FSIQ (2.1 points) scores as compared to WISC-R scores.

McGonagle (1977) examined the comparability of WISC and WISC-R scores for fifty-eight children classified as either mentally retarded, learning disabled, emotionally disturbed or non-special education eligible. The mean test-retest interval for this group
was three years, seven months. Results indicated that VIQ, PIQ and FSIQ scores on the WISC-R were significantly lower than the WISC IQ scores. However, this was not the case for the sample of thirteen mentally retarded children. Although the WISC-R IQs were lower than the WISC IQs for each of the three scales, they were not statistically significant.

Age at Initial Administration

Hamm, Wheeler, McCalum, Herrin, Hunter, and Catoe (1976) examined the comparability of WISC and WISC-R scores for forty-eight mentally retarded children based upon the age of participants. Subjects were divided into two groups. Group I consisted of children aged 9.6 to 10.6. Group II was comprised of children aged 12.6 to 13.6. After an average interval of thirty-nine days, results indicated that no significant mean difference exists based upon age. However, this study examined only children classified as mentally retarded. Therefore, these results can not be generalized to all children in special education.

However, Doppelt and Kaufman (1977) examined the relationship between age and score stability with a homogeneous group of children in special education. Results concluded that the WISC and WISC-R discrepancies were a function of the age of the examinee. The mean FSIQ discrepancy between the two editions of the WISC was about six points for children below age eleven but only two points for children older than eleven. These results were supported by Berry and Sherrets (1975) and Udziela and Barclay (1983) utilizing mentally retarded populations; and by Klinge, Rodziewicz, and Schwartz (1976) utilizing an adolescent psychiatric population.
Ethnicity

Comprised of a sample of twenty volunteer Hispanic children, Munford and Munoz (1980) examined the comparability of WISC and WISC-R scores utilizing a counterbalanced design with a same day test-retest interval. Results indicated high correlations: VIQ=.95, PIQ=.79 and FSIQ=.93. However, the WISC-R yielded lower mean PIQ and FSIQ scores than their WISC counterparts.

Munford (1978) also examined the comparability of the WISC and WISC-R on a sample of twenty African-American children referred for a psychiatric evaluation. Each child was administered the WISC and WISC-R on the same day in a counterbalanced design. Test-retest correlation coefficients were VIQ=.85, PIQ=.82 and FSIQ=.86.

Weiner and Kaufman (1979) examined the comparability of forty-six African-American children ranging in age from seven to ten years. After a test-retest interval of seven weeks, utilizing a counterbalanced design, WISC-R mean scores were significantly lower than WISC scores (VIQ= 7 points, PIQ= 8 points, and FSIQ= 8 points). However, test-retest correlation coefficients were high, VIQ=.90, PIQ=.82 and FSIQ=.90.

Using a sample of 164 children described as African-American, Caucasian, and Hispanic, Swerdlik (1978) examined test-retest scores after a seven to thirty day interval. In all cases, each of the racial groups scored higher on the WISC than on the WISC-R. Correlations for the entire sample were .90 for VIQ, .87 for PIQ and .92 for FSIQ.

Gender

In examining the comparability of WISC and WISC-R Full Scale IQ scores with 101 children enrolled in special education, Covin (1976) reported FSIQ reliability coefficient for females as .94 and for males as .95, indicating no gender differences in IQ constancy.
Summary of WISC/WISC-R Stability Studies

Examination of the literature suggests that extensive research has been conducted investigating the comparability of WISC and WISC-R scores for a variety of populations over variable test-retest intervals. Generally, results suggest that the WISC-R yields lower Verbal IQ, Performance IQ, and Full Scale IQ scores as compared to the WISC. Specifically, children classified as learning disabled, mentally retarded or emotionally disturbed as well as children in “regular education” tend to score significantly lower on the WISC-R across the three IQ measures as compared with the WISC. Investigating the effect of age on the stability of scores, results suggest that children ten years of age and younger experience more score fluctuation than children over ten. Ethnicity and gender of participants does not play a factor in score stability. Therefore, results of this extensive literature review suggest that significant systematic decreases of WISC-R IQ scores as compared to WISC IQ scores exist. Such significant differences in scores between test editions suggest that cognitive assessment materials must be kept up to date. “Reasonably contemporary normative tables are essential for making estimates of a child’s level of intellectual functioning if the goal is to compare him meaningfully with his peers” (Swerdlik, 1977).

Stability of the Wechsler Intelligence Scale for Children- Revised (WISC-R)

Since the publication of the WISC-R in 1974, studies have been conducted to investigate the stability of WISC-R scores over time. Statistical data reported in the WISC-R manual with the standardization sample suggest that the Verbal, Performance and Full Scale IQ scores have reliability coefficients of .94, .90 and .96 respectively across age groups (Wechsler, 1974). Such results were gathered from a subgroup of 303
children from the representative sample who were retested within a three month interval. When comparing the group mean test-retest differences in Verbal, Performance and Full Scale IQ scores, there was a gain of 3.5 points on the Verbal Scale, 9.5 points on the Performance Scale and 7 points on the Full Scale. Because of the short test-retest interval reported in the manual, several studies were conducted to measure the stability of WISC-R IQ scores over an extended period of time, namely the mandated three year re-evaluation period.

Special Education Classification

Non-special education population. Similar results as presented in the WISC-R manual were reported by Tuma and Appelbaum (1980). These researchers examined the degree of stability of the WISC-R when administered to a sample of “normal” children within a six month interval. Each of the forty-five non-special education children were administered the WISC-R twice, with a mean test-retest interval of 5.84 months. Results indicated significant increases of 7.82 points on the PIQ and 4.73 points on the FSIQ. Verbal IQ mean differences were not significant. Test-retest correlation coefficients were .95, .89 and .95 on Verbal, Performance and Full Scale IQ, respectively. As a result, Tuma and Appelbaum (1980) concluded that practice effects are significant for the Performance items among normal children when administered the WISC-R twice.

“Although there is considerable evidence to support the conclusion that WISC-R IQ scores are stable over time in the normal population, one should not infer that the same holds true for exceptional children. Because exceptional children are the ones who are being referred for testing and re-evaluation, evidence concerning the stability of scores in
this group is more pertinent to the question of the need for re-evaluation” (Bauman, 1991, p. 96).

Mentally retarded population. Spitz (1983) examined the constancy of the Full Scale IQ scores for a group of mentally retarded children. Results indicated that after an average interval of two years, there was a significant mean increase of more than three points on the WISC-R FSIQ and a stability correlation of .84.

Learning disabled population. Anderson, Cronin and Kazmierski (1989) examined WISC-R profiles for 113 learning disabled students across two evaluations within a three year interval. Results indicate significant decreases in VIQ by 4.1 points. In addition, correlation coefficients were moderate (VIQ= .55, PIQ= .64 and FSIQ= .58). A similar study conducted by Elliott and Boeve (1987) utilized a three-year test-retest interval with 168 learning disabled children. Results indicated a comparable trend with a two point decrease in VIQ. However in this sample, the subjects’ PIQ increased nearly three points. These results are similar to that of Kaye and Baron (1987) who examined the WISC-R test-retest stability of ninety-nine learning disabled children over a three year period. Results suggested significant VIQ decreases over time, as well as significant PIQ increases over the three year period. With a very small sample of nineteen learning disabled children, Saklofske, Schmidt and Yackulic (1984) discovered the same trend of significant decreases in VIQ and FSIQ between the two administrations conducted three years apart. Haddad, Juliano and Vaughan (1994) studied a large sample of 402 children classified as learning disabled over a similar three year period. Trends were similar to other studies in that there were significant positive changes in PIQ and significant negative changes in VIQ. Pearson correlation coefficients were also high (VIQ= .76, PIQ= .76, FSIQ= .76).
Higher correlation coefficients were reported in a study conducted by Lally, Lloyd and Kulberg (1987) whereby sixty learning disabled children were administered the WISC-R on two separate occasions, approximately 3.23 years apart. No significant VIQ, PIQ, and FSIQ mean differences were determined and coefficients of .73 for VIQ, .89 for PIQ and .82 for FSIQ were found. Schmidt, Kuryliw, Saklofske and Yackulic (1989) also examined the stability of WISC-R scores with a group of children identified as learning disabled. Thirty-six children were evaluated with the WISC-R and subsequently reassessed after a mean interval of 29.4 months. Results indicated significant decreases in VIQ and FSIQ between administrations although no significant changes were noted in the mean PIQ. Weltner-Brunton, Serafica and Friedt (1988) examined a group of fifty-eight children identified as learning disabled. Correlation coefficients appear significantly lower than those reported in the normal population (VIQ=.66, PIQ=.39 and FSIQ=.52). In addition, VIQ scores significantly decreased between administrations approximately three years apart.

Other studies have been conducted that examine the short term stability of WISC-R scores utilizing a learning disabled population. For example, Covin (1977) examined the stability of WISC-R scores for thirty children referred for an evaluation because of learning difficulties after a one day interval. The trends appear similar to the studies reported over a three year period. Verbal IQ scores decreased by .40 points on second administration and PIQ scores increased by 3.74 points. Correlation coefficients were significantly higher than previously discussed studies: VIQ=.83, PIQ=.84 and FSIQ=.85. With a larger sample, Smith and Rogers (1978) examined the stability of the WISC-R with 160 learning disabled children over a six month period. Results reported stability
coefficients as .82 for VIQ, .82 for PIQ, and .79 for FSIQ. After a seven month interval, Smith (1978) examined the stability of WISC-R scores for 161 children enrolled in self-contained classes for learning disabilities. Results appear similar to other studies, in that mean PIQ test-retest change was significantly higher upon retest.

Results of these studies suggest that children classified as learning disabled tend to demonstrate significant instability in WISC-R scores between test-retest administrations. More specifically, it appears that PIQ scores tend to increase over time, whereby VIQ and FSIQ scores tend to decrease over time. Schmidt et al. (1989) suggested three possible reasons why this particular trend exists whereby significant decreases in VIQ and FSIQ occur for the learning disabled population. First, there may be inherent differences between learning disabled children and the general school population. Second, the WISC-R test items expect higher level abstraction and problem-solving for older children. Third, initial reading deficits effect the subsequent development of verbal comprehension skills resulting in a widening gap between normal and learning disabled children.

Learning disabled and mentally retarded populations. Webster (1988) examined the stability of WISC-R scores for a group of eighty-three learning disabled adolescents and a group of seventy-two mentally retarded adolescents all enrolled in self-contained classes. Each child was administered the WISC-R on two separate occasions, approximately three years apart. Results for the mentally retarded group revealed correlation coefficients of .64, .53 and .57 for VIQ, PIQ and FSIQ, respectively. Retest correlations for the learning disabled group indicated greater stability with coefficients of .96, .95 and .99 for VIQ, PIQ and FSIQ. The conclusions may indicate that the stability of WISC-R IQ scores for mentally retarded adolescents is considerably inconsistent and significantly lower than the
stability found for the learning disabled population. In another comparison study with mentally retarded and learning disabled subjects, Whorton (1985) examined eighty-nine mentally retarded students and 221 learning disabled students. Results indicated that no significant differences between administrations were found for VIQ, PIQ or FSIQ for either group. In addition, Pearson correlation coefficients were .59, .60 and .59 for VIQ, PIQ and FSIQ for the mentally retarded group and .65, .58 and .71 for VIQ, PIQ and FSIQ for the learning disabled group.

Many studies have examined the stability of the WISC-R with samples of exceptional children. Thus, by treating them as a homogeneous group, results were reported for both learning disabled and mentally retarded children as one population. Vance, Hankins and Brown (1987) examined the test-retest stability of the three WISC-R administrations of twenty learning disabled children and twelve mentally retarded children, each within a three-year test-retest interval. Results indicated no significant differences in IQ changes for Verbal, Performance or Full Scale IQ scores. Correlation coefficients ranged from .53 to .87 which are lower than those previously reported. However, the researchers attribute this finding to the length of the test-retest interval and the small sample size. Vance, Blixt, Ellis and Debell (1981) reported stability coefficients of .80 for VIQ, .91 for PIQ, and .88 for the FSIQ for thirty mentally retarded and forty-five learning disabled children tested over a two-year test-retest interval. Results indicated no change in mean FSIQ. However, decreases of two points on the VIQ and increases of two points on the PIQ were found.
Full Scale IQ Score at Initial Administration

In order to investigate the effects of initial IQ score on IQ stability, Bauman (1991) divided his 130 elementary school subjects into three groups on the basis of initial FSIQ score. The three groups consisted of children who received an initial FSIQ below 90, which was identified as the "below average" group; 90-110 was identified as the "average" group; and above 110 was classified as the "above average" group. Results indicated that while the mean IQ scores for the entire sample of 130 children significantly declined between testings for VIQ and FSIQ scores, the "above average group" suffered a significantly greater loss on all three subscales than the "below average" IQ group. Bauman (1991) explained that regression toward the mean is not responsible for the loss in scores because this explanation would have predicted that below average groups would make significant gains in IQ scores which did not occur.

Naglieri and Pfeiffer (1983) examined the stability of WISC-R scores with a group of children who scored below 90 on initial administration. Fifty-three children (fifteen classified as mentally retarded, twenty-three classified as borderline intelligence, and fifteen classified as low average) were administered the WISC-R on two separate occasions after a mean interval of two years, ten months. Results indicated no significant mean differences on VIQ, PIQ nor FSIQ scores between administrations. In addition, correlation coefficients suggested that the WISC-R scores have a good test retest reliability over a long period of time (VIQ=.79, PIQ=.75, FSIQ=.85).

Age at Initial Administration

Similar to the studies examining WISC/WISC-R comparability, research has been conducted to determine whether age is related to the degree of change in WISC-R scores.
upon retest. Bauman (1991) examined the relationship of age to IQ stability in children with learning disabilities. With a sample of 130 children, Bauman divided the sample into two groups, below eight at initial administration and over eight at initial administration. The defined reasoning for this distribution is that children below age eight complete Section A of the Coding subtest of the WISC-R, and children over eight complete Section B. After a mean test-retest interval of two years, eight months, results indicated that children under eight experience significant decreases in VIQ, PIQ and FSIQ, suggesting that age is a significant factor in WISC-R change scores over time.

Elliott, Piersel, Witt, Argulewicz, Gutkin, and Galvin (1985) examined the stability coefficients of test-retest VIQ, PIQ, and FSIQ scores. Three hundred and eighty-two children (105 classified as learning disabled, fourteen classified as behavior disordered, and 247 classified as mentally retarded) were divided into three groups according to their age (6.0 years to 9.0 years; 9.1 years to 13.75 years; and 13.75 years to 17.0 years). Results indicated that the stability coefficients of these three age groups did not vary significantly.

Ethnicity

Elliott and Boege (1987) examined the relationship of ethnicity to the three-year stability of the performance of handicapped children on the WISC-R. Subjects consisted of 168 males equally distributed as Caucasian, Mexican-American and African-American and were classified as either learning disabled, mentally retarded or “other.” Results indicated that the variable, ethnicity, did not significantly influence test stability. As a whole, the sample decreased by an average of two points on the mean VIQ and increased by an average of three points on the mean PIQ. “Thus, although statistically significant the influence of three years time on the intelligence test performances of the sampled
handicapped children was pragmatically insignificant” (p. 464). Elliott et al. (1985) also examined the reliability of WISC-R scores with a larger sample consisting of 175 Caucasian, 67 Mexican-American, and 140 African-American children. Of the 382 subjects, 105 were classified as learning disabled, fourteen were classified as behavior disordered and 247 were classified as mentally retarded. Results indicated that with a multiracial group of handicapped children, WISC-R IQ scores were quite stable over a three year test-retest interval.

Gender

Elliott et al. (1985) results suggest that the variable, gender, has a minimal influence on test score stability. However, it was reported that female VIQ resulted in significantly larger stability coefficients than scores obtained by the males in the sample. It was suggested that one possible reason for this finding is that females tend to develop verbal skills earlier and more rapidly than their male counterparts.

Summary of WISC-R Stability Studies

Numerous studies have been conducted that examine the stability of WISC-R scores over time. Because these studies utilized different methodologies, statistical analyses, sample sizes and test-retest intervals, many of the results obtained are inconsistent across investigations. Therefore, while interpreting the results, it is important to understand the manner in which the studies were conducted in order to construct appropriate generalizations. However, the general trend for each variable under investigation is summarized below.

Studies have investigated the relationship of special education classification on WISC-R score stability. Results suggest that children in “regular education” generally experience
stability in WISC-R scores over time. However, significant increases in PIQ and FSIQ scores can be expected. Children classified as mentally retarded experience moderate stability in WISC-R scores across the three IQ measures over time. A significant trend is prevalent for children classified as learning disabled. Generally, Verbal IQ scores decrease and Performance IQ scores increase significantly at retest over a three-year interval.

Examining the effects of initial FSIQ scores on test-retest stability, results suggest that children who receive an initial FSIQ below 90 experience more stability over retest than children who score above 90. Studies investigating the relationship between age and score stability are inconsistent. One study suggests that younger children experience significantly more instability over time than older children while another study suggests that no difference in stability across age group exists. Studies examining the relationship between ethnicity and score stability indicate no difference across ethnic group.

Significant yet pragmatically insignificant differences in gender were found indicating that female VIQ is more stable than male VIQ. Therefore, several variables can be attributed to the instability of WISC-R IQ scores over time.

The Wechsler Intelligence Scale for Children- Third Edition (WISC-III)

The third edition of the Wechsler Intelligence Scale for Children was published in 1991, seventeen years after the second edition. The test is designed for children aged six through sixteen, eleven months. The WISC-III contains thirteen subtests, (six in the Verbal Scale and seven in the Performance Scale). Ten subtests are required and three subtests are supplementary and not computed in the IQ scores.

The WISC-III was standardized with 2,200 children across the United States. One hundred boys and one hundred girls in each of eleven age groups from six years to sixteen
years, eleven months comprised the standardization sample. The sample was stratified on age, ethnicity, geographic region and parental education based upon the 1980 U.S. Census. The WISC-III standardization sampling procedure has been regarded as superior to that of the WISC-R. The WISC-III matched the census quite adequately across variables whereas the WISC-R procedure stratified race by white vs. non-white. WISC-III standardization was also superior to the WISC standardization, which only used white subjects in the standardization sample (Sattler, 1988).

Comparison of WISC-R/WISC-III Scores

Special Education Classification

Non-special education population. The manual of the WISC-III reports a test-retest reliability study conducted with 206 children aged six through sixteen, who were administered the WISC-R and WISC-III in counterbalanced order after a median test interval of 21 days (Wechsler, 1991). Correlation coefficients (VIQ= .90, PIQ= .81 and FSIQ= .89) provided evidence of high test-retest reliability. However, results indicated that mean WISC-III IQ scores significantly decreased when compared with WISC-R mean IQ scores (VIQ= 2 points; PIQ= 7 points; and FSIQ= 5 points).

Several special education populations. The WISC-III manual (Wechsler, 1991) reported a test-retest study consisting of 104 children enrolled in special education, examined as a homogeneous group. Each child was administered the WISC-R and WISC-III. Results suggest that mean VIQ, PIQ and FSIQ scores significantly decreased by five to six points on the WISC-III as compared with the WISC-R.

Graf and Hinton (1994) examined the records of eighty-four children who were administered the WISC-R and WISC-III approximately three years apart. The sample
consisted of children in regular education as well as children classified as learning disabled, behavior disordered, and "unknown." Correlation coefficients for VIQ, PIQ and FSIQ were .82, .81 and .86, respectively.

With a sample of sixty-one children in special education, Bolen, Aichinger, Hall and Webster (1995) examined the stability of WISC-R and WISC-III scores over a retest interval of 2.5 to 3 years. Significantly lower WISC-III IQ scores occur when compared to previously administered WISC-R scores. Mean differences were 5.2 points for VIQ, 9.2 points for PIQ and 8.0 points for FSIQ.

With a group of twenty-seven learning disabled, eighteen emotionally disturbed and twenty-three mentally retarded children, Post (1992) examined the comparability of the WISC-R and WISC-III scores over a mean test-retest interval of three years. Results suggested that WISC-III IQ scores were significantly lower than WISC-R IQ scores for the learning disabled, emotionally disturbed and mentally retarded samples. Mean differences between scores were 6.4 points, 6.5 points and 7.0 points for VIQ, PIQ and FSIQ, respectively.

Learning disabled population. Lyon (1995) examined the records of forty children classified as learning disabled who were administered the WISC-R and WISC-III over a mean period of 2.9 years. Results concluded that WISC-III IQ scores were significantly lower than comparable WISC-R IQ scores. Mean differences were 5.6 points for VIQ, 7.62 for PIQ and 7.05 points for FSIQ. Correlation coefficients were .76 for VIQ, .56 for PIQ and .85 for FSIQ. Dumont and Faro (1993) administered the WISC-III approximately three years subsequent to the administration of the WISC-R to a group of forty-one children classified as learning disabled. Results indicated lower scores of 6.7
points, 8.2 points and 7.7 points for VIQ, PIQ and FSIQ for the WISC-III as compared to previous WISC-R IQ scores. Doll and Boren (1993) examined the performance of seventeen children classified as severely language impaired on the WISC-R and WISC-III. Results indicated that the mean WISC-III VIQ, PIQ, and FSIQ scores were six points lower than respective scores on the WISC-R. Newby, Recht, Cladwell and Schaefer (1993) examined the comparability of WISC-R and WISC-III scores of twenty-six children with dyslexia administered between one and five years apart. Mean VIQ scores were 4.9 points lower and FSIQ scores were 4.8 points lower between the WISC-R and WISC-III. Mean PIQ differences of 3.4 points were not significant.

Mentally retarded population. With a group of ninety-three children classified as mentally retarded, Nagle and Daley (1994) discovered that WISC-III scores were five to eight points lower than comparable WISC-R scores.

Full Scale IQ at Initial Administration

Graf and Hinton (1994) examined the records of eighty-four children who were administered the WISC-R and WISC-III three years apart. In order to examine the effect of initial FSIQ, the entire sample was divided by IQ range. Findings suggest that at the lower IQ range (60-90), the WISC-III yielded higher scores for VIQ, PIQ and FSIQ. However for the 91-105 and the 106-120 subgroups, the WISC-R consistently yielded higher scores across all three IQ measures. These results may be attributed to the phenomena of regression to the mean.

Summary of WISC-R/WISC-III Stability Studies

Extensive research has been conducted to investigate the comparability of WISC-R and WISC-III IQ scores for a variety of populations over variable test-retest intervals.
Generally, results suggest that the WISC-III yields lower Verbal IQ, Performance IQ and Full Scale IQ scores as compared to the previously administered WISC-R. Specifically, children classified as learning disabled, emotionally disturbed or mentally retarded as well as children in "regular education" tend to score significantly lower across the three IQ measures on the WISC-III as compared with the WISC-R. Examining the relationship between initial Full Scale IQ score and score stability, results suggested that children who scored between 60 and 90 on the WISC-R scored significantly higher on the WISC-III for VIQ, PIQ and FSIQ. However, for the children who scored above 90, the WISC-III scores were significantly lower across IQ measures.

WISC-III IQ scores tend to yield significantly lower scores than the WISC-R. This trend is comparable to the studies investigating the stability between WISC and WISC-R scores. Such results suggest that revisions of the Wechsler scales tend to produce lower scores than previous editions.

Stability of the Wechsler Intelligence Scale for Children- Third Edition (WISC-III)

A study described in the WISC-III manual (Wechsler, 1991) addressed the stability of WISC-III scores over time. Three hundred and fifty-three children were tested twice with the WISC-III over a median retest interval of twenty-three days. The stability coefficients of the IQ scores across the age groups were .94, .87 and .94 for Verbal, Performance and Full Scale IQ scores respectively. Thus the WISC-III appears to possess adequate test-retest stability across age groups for a short test-retest interval. However, no studies have investigated the long term stability of WISC-III IQ scores over the mandated three year period.
Statement of Hypotheses

Previous research suggests that IQ scores are generally stable over an extended period of time. Thus, the need for mandated three-year re-evaluations may not be justified for all children in special education. Therefore, the following are hypothesized:

**Hypothesis #1:** Verbal, Performance and Full Scale IQ scores are stable for the populations under investigation.

(a) The correlation coefficients of Verbal, Performance and Full Scale IQ scores for Administration #1 and Administration #2 of the WISC-III will not differ significantly from the median test-retest reliability estimates of the standardization sample.

(b) Group mean differences of test-retest Verbal, Performance and Full Scale IQ scores will not be significant.

(c) There will be significant intra-individual variability between test-retest Verbal, Performance and Full Scale IQ scores.

**Hypothesis #2:** Certain variables may influence test-retest Verbal, Performance and Full Scale IQ scores within specified populations.

**Special education classification**

(a) The learning disabled sample will elicit less stable correlation coefficients for test-retest Verbal, Performance and Full Scale IQ scores than the mentally retarded sample.

(b) The learning disabled sample will experience significant group mean increases in Performance IQ scores and significant group mean decreases in Verbal IQ scores between test-retest administrations.
(c) The mentally retarded population will not experience significant group mean differences in Verbal, Performance and Full Scale IQ scores between test-retest administrations.

Initial Full Scale IQ Score

(d) Students with initial Full Scale IQ scores falling in the mentally deficient or below average category will elicit more stable correlation coefficients for test-retest Verbal, Performance and Full Scale IQ scores than subjects in the above average category.

(e) Group mean differences for test-retest Verbal, Performance and Full Scale IQ scores will not be significant for students with initial Full Scale IQ scores below 90.

(f) Group mean differences for test-retest Verbal, Performance and Full Scale IQ scores will be significant for students with initial Full Scale IQs above 110.

Age at initial administration

(g) Test-retest correlation coefficients for Verbal, Performance and Full Scale IQ scores will be more stable for older children than for children under the age of eight at the time of the initial test administration.

(h) Group mean differences for test-retest Verbal, Performance and Full Scale IQ scores for children under eight will be significantly higher than the group mean differences for children over the age of eight during the initial test administration.

Ethnicity

(i) There will be no significant differences between the correlation coefficients for test-retest Verbal, Performance and Full Scale IQ scores based upon the ethnicity of the child.
(j) Group mean differences for test-retest Verbal, Performance and Full Scale IQ scores will not be significantly different among ethnic groups.

Gender

(k) There will be no significant gender differences between the correlation coefficients of test-retest Verbal, Performance and Full Scale IQ scores.

(l) Group mean differences for test-retest Verbal, Performance and Full Scale IQ scores for males will not be significantly different than the group mean differences for females.
Chapter 2

METHOD

Participants

Longitudinal archival data were collected from the special education files of 592 children who were administered the WISC-III on two separate occasions between September, 1992 and June, 1996 from twenty-seven school districts in a small New England state. Institutional Review Board approval was granted as an exempt review. (See Appendix A). Participants included 396 males (66.9%) and 195 females (32.9%). The gender of one subject was not reported. The sample was comprised of students from five ethnic groups: 492 Caucasians (83.1%), 41 African-Americans (6.9%), 23 Hispanics (3.9%), 8 Asian-Americans (1.4%), and 3 Native-Americans (0.5%). Four students (0.7%) were described as “other” and 21 subjects’ (3.5%) ethnicity was not reported. According to the State’s Department of Education, the following is the Fall, 1996 enrollment summary of all children registered in the public schools in the State: Caucasian (78.2%), Hispanic (10.8%), African-American (7.3%), Asian-American (3.3%) and Native-American (0.5%). These demographics are comparable to those reported in this study.

As a result of the initial administration, students were classified as learning disabled (78.9%), not eligible for services (6.1%), behavior disordered (4.7%), speech and language disordered (4.1%), mentally retarded (2.5%), Attention-Deficit/ Hyperactivity Disorder (1.7%), or Otherwise Health Impaired (1.7%). One student’s classification was not reported. The mean test-retest interval for the entire sample was 35.7 months (SD= 5.7). The mean age of the participants at initial administration was 9.2 years (SD= 2.0) and the mean age at retest was 12.2 years (SD= 2.0). In terms of initial IQ scores, mean
Verbal IQ for the entire sample was 90.7 (SD= 13.98). Mean Performance IQ was 93.56 (SD= 14.55) and mean Full Scale IQ was 91.25 (SD= 13.31).

**Instruments**

The Wechsler Intelligence Scale for Children- Third Edition (WISC-III) is an individually administered cognitive assessment scale for children aged six years through sixteen years, eleven months (Wechsler, 1991). The WISC-III consists of ten subtests and three optional subtests, each measuring a different aspect of intelligence. The child’s performance on these various measures is summarized in three separate IQ scores: a Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ).

Reliability coefficients of the IQ scores as stated in the WISC-III manual over a median 23 day retest interval were .94, .87 and .94 for Verbal, Performance and Full Scale IQ scores, respectively (Wechsler, 1991). Concurrent validity studies reported in the WISC-III manual suggest high correlations between the Full Scale IQ score of the WISC-III and the WISC-R (.89), WAIS-R (.86), and the WPPSI-R (.85). A study reported in the WISC-III manual consisting of 358 children, investigated criterion validity of the WISC-III by means of the correlation between this intelligence measure and school achievement. This study produced significant correlations with a variety of group-administered achievement tests. Correlation coefficients of WISC-III IQ scores and total achievement were .74 for FSIQ, .74 for VIQ, and .57 for PIQ.

**Procedure**

Members of the State organization for school psychologists were asked by mail to participate in the data collection process of the study. The school psychologists were asked to conduct a file review for each child in their school of employment who was
evaluated with the WISC-III on two separate occasions as part of the mandated re-evaluation procedure. A standardized worksheet was provided (See Appendix B). Requested information included demographic information for each student (gender and ethnic background), as well as the age, special education classification, and Verbal, Performance and Full Scale IQ scores for both initial evaluation and re-evaluation. Neither the name nor other identifiable information was recorded on the worksheets which were forwarded to the principal investigator upon completion. Twenty-five school psychologists from twenty-two school districts participated in the data collection process. Each participating school psychologist received a $50.00 stipend which was funded by the State Department of Education, Office of Special Needs (See Appendix B). The principal investigator received permission from the Director of the Office of Special Needs to enter the non-participating school districts to conduct file reviews independently. Information was collected from five school districts by the principal investigator. Thus, data were gathered from twenty-seven out of the thirty-six school districts in the State.
Chapter 3

RESULTS

Hypothesis #1: Stability of IQ Scores for the Entire Sample

In order to examine the long term stability of the Wechsler Intelligence Scale for Children- Third Edition (WISC-III), test-retest Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ) scores were reported for each participant. It was predicted that Verbal, Performance, and Full Scale IQ scores for children in special education would be stable over the three-year mandated re-evaluation period. In examining the stability of scores, three statistical analyses were employed: a correlational method, a test of mean difference, and a test measuring intra-individual variability.

It was predicted that the test-retest correlation coefficients for Verbal, Performance, and Full Scale IQ scores of the WISC-III would not differ significantly from the median test-retest reliability estimates of the standardization sample (Hypothesis 1a). The manual of the WISC-III reports the results of a test-retest reliability study of 353 children over an interval of twenty-three days. Pearson product-moment (PPM) correlation coefficients for test-retest Verbal, Performance, and Full Scale IQ scores were .94, .87, and .94, respectively (Wechsler, 1991). Results of the present study, consisting of 592 subjects over a mean three-year test-retest interval, revealed PPM correlation coefficients for Verbal, Performance, and Full Scale IQ scores as .82, .78, and .84, respectively. To compare stability coefficients statistically across Verbal, Performance, and Full Scale IQ scores, each PPM correlation coefficient was transformed to a Fisher Z, and the difference (dz) between the transformed z’s was calculated for the desired pairs of correlations (Yaremko, Harari, Harrison, & Lynn, 1982). The resulting comparisons revealed
statistically significant differences in test-retest stability coefficients for Verbal IQ, $z = -8.61$, $p < .05$; Performance IQ, $z = -4.26$, $p < .05$; and Full Scale IQ, $z = -7.63$, $p < .05$; scores between the Wechsler stability study and this current investigation.

It was predicted that group mean differences between test-retest Verbal, Performance, and Full Scale IQ scores would not be significant over the three-year re-evaluation period (Hypothesis 1b). Five hundred and ninety-two cases reported test-retest Verbal IQ and Performance IQ scores. Five hundred and eighty cases reported test-retest Full Scale IQ scores. (See Table 1 for means and standard deviations).

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verbal IQ</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 1</td>
<td>90.70</td>
<td>13.98</td>
</tr>
<tr>
<td>Time 2</td>
<td>90.31</td>
<td>14.01</td>
</tr>
<tr>
<td><strong>Performance IQ</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 1</td>
<td>93.56</td>
<td>14.55</td>
</tr>
<tr>
<td>Time 2</td>
<td>93.54</td>
<td>14.89</td>
</tr>
<tr>
<td><strong>Full Scale IQ</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 1</td>
<td>91.37</td>
<td>13.29</td>
</tr>
<tr>
<td>Time 2</td>
<td>91.05</td>
<td>13.64</td>
</tr>
</tbody>
</table>

In order to investigate whether significant mean differences exist between test-retest VIQ, PIQ and FSIQ scores for the entire sample, two-tailed $t$-tests were conducted. No significant differences were found between test-retest Verbal IQ scores, $t(591) = 1.12$, $p >$
.05; Performance IQ scores, \( t(591) = .05, p > .05 \); nor Full Scale IQ scores, \( t(579) = 1.01, p > .05 \).

It was predicted that intra-individual variability between test-retest Verbal, Performance, and Full Scale IQ scores for the entire sample would exist (Hypothesis 1c). This hypothesis was not confirmed. The Verbal, Performance and Full Scale IQ mean differences between the first and second administrations of the WISC-III were not significant. Verbal IQ scores increased by a mean of .39 points between administrations, \( t(591) = 1.12, p > .05 \); Performance IQ scores increased by a mean of .05 points between administrations, \( t(591) = .047, p > .05 \); and Full Scale IQ scores increased by a mean of .32 points between administrations, \( t(579) = 1.01, p > .05 \). Therefore, intra-individual variability over the three-year test-retest interval for Verbal IQ, Performance IQ, and Full Scale IQ for this sample of 592 children in special education was not significant.

Results of the three statistical analyses employed (test of correlations, test of mean differences and test of individual variability) suggest that Verbal IQ, Performance IQ and Full Scale IQ scores are stable for this sample of children in special education who were administered the WISC-III twice over the three-year mandated re-evaluation period.

Hypothesis #2: Stability of IQ Scores Within Specified Populations

The above mentioned results indicated that WISC-III Verbal, Performance and Full Scale IQ scores are stable over the three-year re-evaluation period. However, in order to determine whether specified populations will elicit less stable scores, several subgroups were examined (special education classification, Full Scale IQ score at initial administration, age of the participants at initial administration, as well as the ethnicity and gender of the participants).
Special Education Classification

The special education files of 592 children were examined to investigate the stability of WISC-III IQ scores over a three-year re-evaluation period. As a result of the initial administration, these children were classified according to State regulations as learning disabled (n= 467), mentally retarded (n= 15), behavior disordered (n= 28), not eligible for services (n= 36), Attention-Deficit/Hyperactivity Disorder (n= 10), Otherwise Health Impaired (n= 10), and speech/language disordered (n= 24). The classifications of two children were not reported. (See Table 2 for means, standard deviations, correlation coefficients, and tests of mean difference for VIQ, PIQ and FSIQ for each subgroup).

Results of the tests of mean difference suggested that three-year test-retest VIQ, PIQ, and FSIQ scores were not significant for children diagnosed with the following classifications: learning disabled, mentally retarded, not eligible for services, ADHD, Otherwise Health Impaired, and speech/language disordered. A significant difference was found between test-retest Performance IQ scores for the behavior disordered sample, t(27)= 2.14, p < .05. PIQ scores for this sample decreased by a mean of four points between WISC-III administrations. No significant differences were found between test-retest VIQ and FSIQ scores for this sample.

Correlation coefficients were calculated for test-retest VIQ, PIQ and FSIQ scores for each special education classification under consideration. Pearson product-moment (PPM) correlation coefficients were calculated for the 'learning disabled' and 'not eligible for services' samples. Because the other subgroups consisted of fewer than thirty pairs of measurements, a "rank-difference correlation coefficient," or Spearman's rho was calculated (Cohen, Montague, Nathanson, & Swerdlik, 1988).
Table 2

Means, Standard Deviations, Correlation Coefficients and Tests of Mean Differences (t) Between Time 1 and Time 2 Based on Initial Special Education Classification

<table>
<thead>
<tr>
<th>Initial Classification</th>
<th>Time 1</th>
<th></th>
<th>Time 2</th>
<th></th>
<th>n</th>
<th>Corr. Coeff.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning Disabled</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>90.89</td>
<td>13.37</td>
<td>90.58</td>
<td>13.28</td>
<td>467</td>
<td>.80***</td>
<td>.76</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>94.89</td>
<td>13.59</td>
<td>95.07</td>
<td>14.03</td>
<td>467</td>
<td>.75***</td>
<td>-.34</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>92.19</td>
<td>12.24</td>
<td>92.00</td>
<td>12.59</td>
<td>456</td>
<td>.82***</td>
<td>.51</td>
</tr>
<tr>
<td>Mentally retarded</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>65.87</td>
<td>7.44</td>
<td>65.13</td>
<td>9.15</td>
<td>15</td>
<td>.34</td>
<td>.30</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>66.60</td>
<td>10.58</td>
<td>64.13</td>
<td>9.67</td>
<td>15</td>
<td>.75***</td>
<td>1.42</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>62.86</td>
<td>9.05</td>
<td>61.64</td>
<td>9.20</td>
<td>14</td>
<td>.59*</td>
<td>.62</td>
</tr>
<tr>
<td>Behavior disordered</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>93.29</td>
<td>15.10</td>
<td>92.57</td>
<td>16.50</td>
<td>28</td>
<td>.81***</td>
<td>.34</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>94.75</td>
<td>16.07</td>
<td>90.64</td>
<td>18.37</td>
<td>28</td>
<td>.79***</td>
<td>2.14*</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>93.57</td>
<td>14.14</td>
<td>90.86</td>
<td>16.76</td>
<td>28</td>
<td>.89***</td>
<td>1.57</td>
</tr>
<tr>
<td>Not eligible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>94.17</td>
<td>12.58</td>
<td>94.39</td>
<td>10.90</td>
<td>36</td>
<td>.81***</td>
<td>-.18</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>93.75</td>
<td>12.62</td>
<td>93.36</td>
<td>13.23</td>
<td>36</td>
<td>.70***</td>
<td>-1.56</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>92.92</td>
<td>12.37</td>
<td>94.61</td>
<td>11.40</td>
<td>36</td>
<td>.76***</td>
<td>-.123</td>
</tr>
<tr>
<td>ADHD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>96.30</td>
<td>9.24</td>
<td>98.90</td>
<td>11.75</td>
<td>10</td>
<td>.61*</td>
<td>-1.00</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>98.90</td>
<td>18.33</td>
<td>94.80</td>
<td>19.32</td>
<td>10</td>
<td>.77**</td>
<td>1.37</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>97.90</td>
<td>12.74</td>
<td>95.70</td>
<td>15.80</td>
<td>10</td>
<td>.64*</td>
<td>.84</td>
</tr>
<tr>
<td>Otherwise Health Impaired</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>95.60</td>
<td>7.15</td>
<td>96.40</td>
<td>19.52</td>
<td>10</td>
<td>.88***</td>
<td>-.24</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>89.30</td>
<td>17.88</td>
<td>87.70</td>
<td>14.19</td>
<td>10</td>
<td>.90***</td>
<td>.53</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>91.70</td>
<td>18.95</td>
<td>91.60</td>
<td>16.24</td>
<td>10</td>
<td>.88***</td>
<td>.03</td>
</tr>
<tr>
<td>Speech /Language Disordered</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>84.67</td>
<td>11.96</td>
<td>84.21</td>
<td>11.85</td>
<td>24</td>
<td>.80***</td>
<td>.42</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>88.00</td>
<td>10.59</td>
<td>90.63</td>
<td>8.18</td>
<td>24</td>
<td>.72***</td>
<td>-1.62</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>85.00</td>
<td>10.13</td>
<td>86.08</td>
<td>9.17</td>
<td>24</td>
<td>.83***</td>
<td>-1.01</td>
</tr>
</tbody>
</table>

***p < .001. **p < .01. *p < .05.
Results confirmed significant test-retest correlation coefficients for VIQ, PIQ and FSIQ scores for each special education classification under consideration except for Verbal IQ in the mentally retarded sample; for which the correlation coefficient of .34 was not significant. (See Table 2). Analyses were conducted to determine if significant intra-individual variability existed for test-retest Verbal IQ, Performance IQ and Full Scale IQ scores based upon the special education classification of the child. Significant intra-individual variability did not exist between test-retest VIQ, PIQ nor FSIQ scores for the learning disabled, mentally retarded, not eligible for services, ADHD, Otherwise Health Impaired and speech/language disordered subgroups.

For the learning disabled sample, VIQ scores increased by a mean of .47 points, \( t(466) = 1.20, p > .05 \); PIQ scores decreased by a mean of .09 points, \( t(466) = -1.92, p > .05 \); and FSIQ scores increased by a mean of .33 points between administrations, \( t(455) = 0.941, p > .05 \). For the mentally retarded sample, VIQ scores increased by a mean of .73 points, \( t(14) = 0.299, p > .05 \); PIQ scores increased by a mean of 2.47 points, \( t(14) = 1.418, p > .05 \); and FSIQ scores increased by a mean of 1.21 points between administrations, \( t(13) = 0.622, p > .05 \). For the 'not eligible for services' subgroup, VIQ scores decreased by a mean of .22 points, \( t(35) = -0.179, p > .05 \); PIQ scores decreased by a mean of 2.61 points, \( t(35) = -1.565, p > .05 \); and FSIQ scores decreased by a mean of 1.69 points between administrations, \( t(35) = -1.226, p > .05 \). For the ADHD subgroup, VIQ scores decreased by a mean of 2.60 points, \( t(9) = -0.997, p > .05 \); PIQ scores increased by a mean of 4.10 points, \( t(9) = 1.371, p > .05 \); and FSIQ scores increased by a mean of 2.20 points between administrations, \( t(9) = 0.838, p > .05 \). For the Otherwise Health Impaired subgroup, VIQ scores decreased by a mean of .80 points, \( t(9) = -0.244, p > .05 \); PIQ scores increased by a
mean of 1.60 points, $t(9) = .532, p > .05$; and FSIQ scores increased by a mean of .10 points between administrations, $t(9) = .032, p > .05$. For the speech/language disordered subgroup, VIQ scores increased by a mean of .46 points, $t(23) = .424, p > .05$; PIQ scores decreased by a mean of 2.63 points, $t(23) = -1.618, p > .05$; and FSIQ scores decreased by a mean of 1.08 points between administrations, $t(23) = -1.009, p > .05$.

The only subgroup that contained significant intra-individual variability was the behavior disordered subgroup. Performance IQ scores increased by a mean of 4.11 points between administrations. Such an increase in mean scores is significant, $t(27) = 2.138, p < .05$. Test-retest VIQ and FSIQ score differences were not significant for this sample. VIQ for the behavior disordered subgroup increased by a mean of .71 points, $t(27) = .338, p > .05$; and FSIQ scores increased by a mean of 2.71 points between administrations, $t(27) = 1.572, p > .05$.

It was predicted that the learning disabled sample would elicit less stable correlation coefficients for test-retest Verbal, Performance and Full Scale IQ scores than the mentally retarded sample (Hypothesis 2a). (Refer to Table 2 for test-retest correlation coefficients). Because Fisher's Z test only examines the significance of the difference between Pearson product moment (PPM) correlation coefficients, PPM coefficients were calculated for the learning disabled and mentally retarded groups. Pearson product-moment correlation coefficient for test-retest Verbal IQ scores for the learning disabled sample was .80 and .36 for the mentally retarded population. Utilizing a one-tailed Fisher's Z, $z = 2.47, p < .05$, results indicated a significant difference in correlation coefficients in test-retest Verbal IQ for the learning disabled and mentally retarded samples. This significant difference indicated opposite results than was hypothesized,
whereby the correlation coefficients for VIQ were more stable for the learning disabled sample than for the mentally retarded sample. PPM correlation coefficient for test-retest Performance IQ scores for the learning disabled sample was .74 and .78 for the mentally retarded sample, indicating no significant difference between test-retest correlation coefficients for PIQ, $z = -.32, p > .05$. Test-retest PPM correlation coefficients for Full Scale IQ scores for the learning disabled population was .82 and .68 for the mentally retarded population. One-tailed Fisher’s Z test suggested no significant difference between correlation coefficients of test-retest FSIQ for the learning disabled and mentally retarded samples, $z = 1.07, p > .05$.

It was predicted that the learning disabled sample would experience significant group mean increases in Performance IQ scores and significant decreases in Verbal IQ scores between test-retest administrations (Hypothesis 2b). (Refer to Table 2 for means, standard deviations, and t-values). Because the direction of the difference was hypothesized, one-tailed within-group t-tests were utilized. No significant differences were found between the mean Verbal IQ scores, $t(424) = .76, p > .05$; Performance IQ scores, $t(424) = .37, p > .05$; nor Full Scale IQ scores between administrations, $t(413) = .51, p > .05$, for the learning disabled sample. Therefore, VIQ, PIQ and FSIQ scores are stable over a three-year test-retest interval for the learning disabled population.

It was predicted that the mentally retarded sample would not experience significant group mean differences in Verbal, Performance, and Full Scale IQs between test-retest administrations (Hypothesis 2c). (Refer to Table 2 for means, standard deviations and t-values). Utilizing two-tailed t-tests, no significant differences were discovered between test-retest Verbal IQ scores, $t(14) = .30, p > .05$; Performance IQ scores, $t(14) = 1.42, p > .05$. 

50
Therefore, VIQ, PIQ and FSIQ scores are stable over a three-year test-retest interval for a mentally retarded population.

Initial Full Scale IQ Score

Subjects were divided into three groups based on their Full Scale IQ score at initial administration. Subjects scoring below 90 were categorized as Group 1. Subjects scoring between 90 and 109 were categorized as Group 2, and subjects scoring 110 or above were categorized as Group 3. (See Table 3 for means, standard deviations, PPM correlation coefficients and t-values).

It was predicted that students with an initial Full Scale IQ score below 90 would have more stable PPM correlation coefficients than subjects scoring above 109 (Hypothesis 2d). (Refer to Table 3). The difference between test-retest PPM correlation coefficients for VIQ (Group 1, r= .69; Group 3, r= .51) utilizing a one-tailed Fisher’s Z test was significant, z = 1.15, p < .05. Similarly, the difference between test-retest PPM correlation coefficients for PIQ (Group 1, r= .70; Group 3, r= .40) was also significant, z = 2.72, p < .05. In addition, the difference between test-retest PPM correlation coefficients for FSIQ (Group 1, r= .73; Group 3 r= .52) was significant, z = 2.15, p < .05. Such results suggested that subjects with initial FSIQ below 90 have more stable test-retest VIQ, PIQ and FSIQ scores than subjects with initial FSIQ above 109.

It was predicted that group mean differences for test-retest VIQ, PIQ and FSIQ scores would not be significant for students with initial FSIQ scores below 90 (Hypothesis 2e). Results did not confirm this hypothesis. Utilizing two-tailed t-tests, results suggested that VIQ, PIQ and FSIQ scores significantly increase over a three-year test-retest interval
for students with initial FSIQ scores below 90. The mean VIQ score was significantly higher at retest, \(t(256) = -2.44, p < .05\); mean PIQ score was significantly higher at retest \(t(256) = -2.89, p < .05\); and mean FSIQ score was significantly higher at retest, \(t(251) = -3.26, p < .05\). Therefore, results suggested that group mean differences exist between test-retest VIQ, PIQ and FSIQ for students with initial FSIQ below 90. These students will score significantly higher on the second administration approximately three years later.

Table 3

Means, Standard Deviations, Correlation Coefficients, and Tests of Mean Differences (t) Between Time 1 and Time 2 Based on Full Scale IQ Score at Initial Administration

<table>
<thead>
<tr>
<th>Initial FSIQ Score</th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td><strong>Group 1 (&lt;90)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>79.81</td>
<td>9.72</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>82.40</td>
<td>10.40</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>79.42</td>
<td>8.14</td>
</tr>
<tr>
<td><strong>Group 2 (90-109)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>97.03</td>
<td>8.74</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>99.84</td>
<td>9.15</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>97.99</td>
<td>5.47</td>
</tr>
<tr>
<td><strong>Group 3 (&gt;109)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>112.55</td>
<td>7.90</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>116.87</td>
<td>8.63</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>115.83</td>
<td>4.26</td>
</tr>
</tbody>
</table>

***p < .001. **p < .01. *p < .05.

It was predicted that group mean differences for test-retest VIQ, PIQ and FSIQ scores would be significant for students with initial FSIQ scores above 110 (Hypothesis
Utilizing two-tailed t-tests, results suggested that VIQ, PIQ and FSIQ scores decrease over a three-year test-retest interval for individuals scoring an initial FSIQ above 110. The mean VIQ score was significantly lower at retest, \( t(46) = 2.42, p < .05 \); the mean PIQ score was significantly lower at retest, \( t(46) = 3.93, p < .05 \); and the mean FSIQ score was significantly lower at retest, \( t(46) = 4.56, p < .05 \). Therefore, results indicated that children scoring an FSIQ above 109 score significantly lower on VIQ, PIQ and FSIQ on a second administration three years later.

Students who received a FSIQ score between 90 and 109 at the initial administration scored significantly lower on VIQ at the second administration three years later, \( t(280) = 2.74, p < .05 \). The same subgroup also scored significantly lower on FSIQ on the second administration, \( t(280) = 2.61, p < .05 \). No significant difference was found between test-retest PIQ for this group, \( t(280) = 1.10, p > .05 \).

Analyses were conducted to determine if significant intra-individual variability existed for VIQ, PIQ and FSIQ scores based upon the initial Full Scale IQ score. For the group who received an initial FSIQ score below 90, significant intra-individual variability existed between test-retest VIQ, PIQ and FSIQ scores. VIQ scores decreased by a mean of 1.24 points, \( t(256) = -2.436, p < .05 \); PIQ scores decreased by a mean of 1.65 points, \( t(256) = -2.890, p < .05 \); and FSIQ scores decreased by a mean of 1.51 points between administrations, \( t(251) = -3.260, p < .05 \). For the group who received an initial FSIQ score between 90 and 109, significant intra-individual variability existed for test-retest VIQ and FSIQ scores. VIQ scores increased by a mean of 1.38 points, \( t(280) = 2.736, p < .05 \); and FSIQ scores increased by a mean of 1.12 points, \( t(280) = 2.611, p < .05 \). No significant intra-individual variability existed for test-retest PIQ scores for this group. PIQ scores
increased by a mean of .62 points between administrations, \( t(280)= 1.104, p > .05 \). For the group who received an initial FSIQ score above 109, significant intra-individual variability existed across the three IQ measures. VIQ scores increased by a mean of 3.13 points, \( t(46)= 2.419, p < .05 \); PIQ scores increased by a mean of 6.36 points, \( t(46)= 3.934, p < .05 \); and FSIQ scores increased by a mean of 5.32 points between administrations, \( t(46)= 4.646, p < .05 \).

Therefore, significant intra-individual variability existed for test-retest VIQ, PIQ and FSIQ scores for the samples who scored below 90 and above 109 on FSIQ on initial administration. For the sample who scored in the average range (90-109), significant intra-individual variability existed for VIQ and FSIQ.

**Age at Initial Administration**

Subjects were divided into two groups: children under the age of eight (Group 1) and children eight years of age and older (Group 2) at initial administration. (See Table 4 for means, standard deviations, and correlation coefficients).

It was predicted that test-retest PPM correlation coefficients for VIQ, PIQ and FSIQ would be more stable for children over the age of eight than for children under the age of eight at the time of the initial administration (Hypothesis 2g). (See Table 4). Test-retest PPM correlation coefficient for VIQ for Group 1 was .84, and .81 for Group 2. One-tailed Fisher’s Z test suggested that no significant difference exists between test-retest correlations for VIQ based on age at initial administration, \( z=1.06, p > .05 \). Test-retest PPM correlation coefficient for FSIQ for Group 1 was .84 and .84 for Group 2, indicating no difference. On the other hand, test-retest PPM correlation differences for PIQ were significant (Group 1, \( r=.71 \); Group 2, \( r=.81 \), \( z=-2.70, p < .05 \), suggesting that children
eight and older had more stable PIQ scores than children below age eight over the three-year test-retest interval.

Table 4

Means, Standard Deviations and Correlation Coefficients For Time 1 and Time 2 Based on Age of Participants at Initial Administration

<table>
<thead>
<tr>
<th>Age at Time 1</th>
<th>Time 1</th>
<th>Time 2</th>
<th>n</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Group 1 (&lt;8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>92.90</td>
<td>14.70</td>
<td>93.59</td>
<td>13.77</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>96.13</td>
<td>15.21</td>
<td>94.32</td>
<td>14.53</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>93.59</td>
<td>13.76</td>
<td>93.05</td>
<td>13.29</td>
</tr>
<tr>
<td>Group 2 (&gt;8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>89.61</td>
<td>13.49</td>
<td>88.77</td>
<td>13.89</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>92.31</td>
<td>14.07</td>
<td>93.13</td>
<td>15.04</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>90.26</td>
<td>12.90</td>
<td>90.07</td>
<td>13.71</td>
</tr>
</tbody>
</table>

***p < .001, **p < .01, *p < .05.

It was predicted that group mean differences for test-retest Verbal, Performance and Full Scale IQ scores for children under the age of eight would be significantly higher than the group mean differences for children eight years of age and older (Hypothesis 2h).

(Refer to Table 4 for means and standard deviations). A multivariate repeated measures analysis of variance (MANOVA) was conducted whereby the independent variables were age and time, and the dependent variables were the three IQ measures. Utilizing Wilk's test of significance, a significant interaction effect suggested that there was significant differential change over time by age group, F(3,575) = 5.62, p < .05. Three separate 2x2
univariate ANOVAs were utilized to determine the age group x time interaction for each IQ measure. Statistically significant group mean differences were found between test-retest VIQ scores based upon the age of the child at the initial administration, $F(1,577)=4.01$, $p < .05$. Group mean VIQ scores for Group 1 improved by .69 points between administrations, whereas the group mean VIQ score for Group 2 decreased by .84 points between administrations. Similarly, statistically significant group mean differences existed between test-retest PIQ scores based upon the age at initial administration, $F(1,577)=9.57$, $p < .05$. Results suggested that group mean PIQ scores for Group 1 decreased by 1.81 points between administrations, whereas the group mean PIQ score for Group 2 increased by .82 points between administrations. No significant difference was found between the group mean FSIQ difference between test and retest, $F(577)=.277$, $p > .05$.

Analyses were conducted in order to determine whether intra-individual variability existed for test-retest VIQ, PIQ and FSIQ scores based upon the age of the child at initial administration. For the subgroup of children who were under the age of eight at the time of the initial administration, statistically significant intra-individual variability existed for test-retest PIQ scores. PIQ scores increased by a mean of 1.80 points between administrations, $t(187)=2.197$, $p < .05$. Intra-individual variability was not significant for test-retest VIQ and FSIQ scores for this subgroup. VIQ scores decreased by a mean of .69 points, $t(187)=1.156$, $p > .05$; and FSIQ scores increased by a mean of .54 points between administrations, $t(184)=.961$, $p > .05$. For the subgroup of children who were eight years of age and older at the time of the initial administration, intra-individual variability was significant for test-retest VIQ. VIQ scores increased by a mean of .84 points between administrations, $t(402)=1.988$, $p < .05$. Intra-individual variability was not
significant for test-retest PIQ nor FSIQ scores for this subgroup. PIQ scores decreased by a mean of .82 points, \( t(402) = -1.839, p > .05 \); and FSIQ scores increased by a mean of .19 points between administrations, \( t(393) = .488, p > .05 \).

**Ethnicity of Participants**

It was predicted that there would be no significant difference between the PPM correlation coefficients from test-retest Verbal, Performance and Full Scale IQ scores based upon the ethnicity of the child (Hypothesis 2i). In addition, it was predicted that group mean differences for test-retest Verbal, Performance and Full Scale IQ scores would not be statistically significant among ethnic groups (Hypothesis 2j). Because of the small percentage of participants that were reported as other than Caucasian (17.2%), the analysis divided ethnicity into two categories “Caucasian” and “non-Caucasian.” (See Table 5 for means, standard deviations, and PPM correlation coefficients).

A two-tailed Fisher’s Z test was utilized for each pair of test-retest PPM correlation coefficients for the three IQ measures for Caucasians and non-Caucasians. Results suggested that no significant differences between correlation coefficients for Verbal IQ scores, \( z = -.73, p > .05 \); Performance IQ scores, \( z = -1.36, p > .05 \); nor Full Scale IQ scores, \( z = -.54, p > .05 \) exist based on the ethnicity of the subjects.

In order to determine whether significant group mean differences exist for test-retest VIQ, PIQ and FSIQ scores between the Caucasian and non-Caucasian group, a multivariate repeated measures analysis of variance (MANOVA) was conducted. Independent variables were ethnicity and time; and the dependent variables were the three IQ measures. Utilizing Wilk’s test of significance, results suggested that no significant interaction exist between Caucasians and non-Caucasians in the amount of change.
between administrations across IQ measures $F(3,555)= 1.08, p > .05$. Therefore, no significant test-retest differences were found between the amount of change across VIQ, PIQ and FSIQ between Caucasians and non-Caucasians.

Table 5

Means, Standard Deviations and Correlation Coefficients For Time 1 and Time 2 Based on Ethnicity of Participants

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Time 1</th>
<th>Time 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>n</td>
<td>r</td>
</tr>
<tr>
<td><strong>Caucasian</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>91.86</td>
<td>13.44</td>
<td>91.41</td>
<td>13.47</td>
<td>492</td>
<td>.80***</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>94.31</td>
<td>14.05</td>
<td>94.41</td>
<td>14.76</td>
<td>492</td>
<td>.77***</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>92.38</td>
<td>12.74</td>
<td>92.05</td>
<td>13.25</td>
<td>482</td>
<td>.83***</td>
</tr>
<tr>
<td><strong>Non-Caucasian</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>81.87</td>
<td>13.90</td>
<td>81.49</td>
<td>14.54</td>
<td>79</td>
<td>.83***</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>89.38</td>
<td>16.81</td>
<td>87.47</td>
<td>15.22</td>
<td>79</td>
<td>.83***</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>84.39</td>
<td>14.39</td>
<td>83.30</td>
<td>14.17</td>
<td>79</td>
<td>.85***</td>
</tr>
</tbody>
</table>

***p < .001. **p < .01. *p < .05.

Analyses were conducted in order to determine whether intra-individual variability existed between test-retest VIQ, PIQ and FSIQ scores based upon the ethnicity of the participants. Results indicated that intra-individual variability was not significant between test-retest scores across the three IQ measures for each subgroup. For the Caucasian group, VIQ scores increased by a mean of .45 points, $t(491)= 1.178, p > .05$; PIQ scores decreased by a mean of .10 points, $t(491)= -.221, p > .05$; and FSIQ scores increased by .34 points, $t(481)= .983, p > .05$, between administrations. For the non-Caucasian group,
VIQ scores increased by a mean of .38 points, *t*(78) = .411, *p* > .05; PIQ scores increased by a mean of 1.91 points, *t*(78) = 1.779, *p* > .05; and FSIQ scores increased by a mean of 1.09 points, *t*(76) = 1.237, *p* > .05, between administrations.

**Gender of Participants**

It was predicted that gender differences would not exist in the stability of IQ scores over time. More specifically, it was predicted that there would be no significant differences between the correlation coefficients of test-retest Verbal, Performance and Full Scale IQ scores based upon the gender of the child (Hypothesis 2k). It was also predicted that group mean differences for test-retest Verbal, Performance and Full Scale IQ scores for males would not be significantly different than the group mean differences for females (Hypothesis 2l). (See Table 6 for means, standard deviations, and PPM correlation coefficients).

PPM correlation coefficients for test-retest Verbal IQ scores for males were .81 and .83 for females. Utilizing a two-tailed Fisher’s Z test, results indicated that no significant difference exists between correlation coefficients of test-retest Verbal IQ scores based on the gender of the subjects, *z* = .69, *p* > .05. PPM correlation coefficients for test-retest Full Scale IQ scores were .83 for males and .86 for females. Utilizing a two-tailed Fisher’s Z test, no significant difference was reported between the correlation coefficients of test-retest Full Scale IQ scores based on gender, *z* = -1.19, *p* > .05. Test-retest PPM correlation coefficients for Performance IQ were .75 for males and .82 for females. Utilizing a two-tailed Fisher’s Z test, results indicated a significant difference between correlation coefficients of test-retest Performance IQ based on gender, *z* = -2.09, *p* < .05.
Analysis of the stability coefficients revealed that females' PIQ scores are significantly more stable than male Performance IQ scores over a three-year test-retest interval.

Table 6
Means, Standard Deviations, and Correlation Coefficients For Time 1 and Time 2 Based on the Gender of the Participants

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
<th>n</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Males</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>91.98</td>
<td>14.27</td>
<td>91.63</td>
<td>14.40</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>94.49</td>
<td>14.36</td>
<td>94.63</td>
<td>14.35</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>92.64</td>
<td>13.26</td>
<td>92.44</td>
<td>13.27</td>
</tr>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>87.98</td>
<td>12.93</td>
<td>87.53</td>
<td>12.75</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>91.67</td>
<td>14.83</td>
<td>91.31</td>
<td>15.77</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>88.76</td>
<td>13.01</td>
<td>88.20</td>
<td>13.96</td>
</tr>
</tbody>
</table>

***p < .001. **p < .01. *p < .05.

To determine if the group mean difference between administrations for Verbal, Performance, and Full Scale IQ scores differed for males and females, a multivariate repeated measures analysis of variance (MANOVA) was utilized. (Refer to Table 6 for means and standard deviations). Independent variables were gender and time; and dependent variables were the three IQ measures. Utilizing Wilk's test of significance, results suggested that no significant interaction exists between males and females in the amount of change from Time 1 to Time 2 across IQ measures, F(3,575)= .16, p > .05.
Analyses were conducted in order to determine whether significant intra-individual variability existed between test-retest scores for VIQ, PIQ and FSIQ based upon the gender of the participants. Results suggested that intra-individual variability was not significant for males nor females across the three IQ measures. For the male subgroup, VIQ scores increased by a mean of .35 points, $t(395)= .783$, $p > .05$; PIQ scores decreased by a mean of .14 points, $t(395)= .543$, $p > .05$; and FSIQ scores increased by a mean of .20 points, $t(385)= .506$, $p > .05$ between administrations. For the female subgroup, VIQ scores increased by a mean of .45 points, $t(194)= .835$, $p > .05$; PIQ scores increased by a mean of .35 points, $t(194)= .543$, $p > .05$; and FSIQ scores increased by a mean of .56 points, $t(192)= 1.079$, $p > .05$ between administrations.
Chapter 4

DISCUSSION

The purpose of the current study was to investigate the long term stability of Verbal, Performance and Full Scale IQ scores for a sample of exceptional children evaluated with the Wechsler Intelligence Scale for Children- Third Edition over a three-year interval. Several variables (special education classification, Full Scale IQ score at initial administration, age of the participant at initial administration, ethnicity and gender of the participants) were examined to detect which variables, if any, would influence the stability of test scores over the mandated three-year re-evaluation period.

Stability of IQ Scores for the Entire Sample

Results of the current investigation indicated strong reliability between test-retest Verbal, Performance, and Full Scale IQ scores over a three-year interval for 592 exceptional children. However, the Pearson product-moment (PPM) correlation coefficients were significantly less than the reported PPM correlation coefficients of the standardization sample, consisting of 353 non-exceptional children, reported in the WISC-III manual. There may be two reasons for the discrepancy in reliability coefficients. First, the length of the test-retest interval may be a factor. The mean test-retest interval of the current investigation was three years, whereas the test-retest interval of the standardization sample was a median of twenty-three days. Because stability coefficients tend to decrease over time, such discrepancy in correlation coefficients may simply be due to the differential time interval between the Wechsler study and the current study. A second explanation for the discrepancy in reliability coefficients between this investigation and the Wechsler reliability study is that the current study was comprised of children in
special education, whereas the Wechsler sample consisted of a random selection of the entire school-aged population, mainly children not receiving special education services. Therefore, comparisons are difficult to interpret because the samples themselves are intrinsically different.

Based on previous research investigating the long term stability of the WISC-R, it was predicted that no significant mean differences would exist between test-retest VIQ, PIQ and FSIQ scores. Results confirmed this hypothesis, whereby no significant mean differences were found between test-retest VIQ, PIQ and FSIQ scores for this sample of 592 children treated as a homogeneous group. Results of this investigation support previous studies examining the stability of IQ scores within a homogeneous special education population (Whorton, 1985; Vance, Hankins & Brown, 1987).

Previous research investigating the stability of the WISC-R with a homogeneous group of children in special education suggest that significant intra-individual variability exist across the three IQ measures (Elliott & Boeve, 1987; Truscott, Narrett & Smith, 1994; Vance, Blixt, Ellis & Debell, 1981; Webster, 1988). Results of this investigation do not confirm these findings. No significant intra-individual variability existed for this sample of children across the three IQ measures. Such results indicated that for this sample of 592 children in special education, treated as a homogeneous group, WISC-III VIQ, PIQ and FSIQ scores remain stable over the three-year mandated test-retest re-evaluation period.

Stability of IQ Scores Within Specified Populations

Results suggested that for this sample of 592 children in special education, examined as a homogeneous group, WISC-III Verbal, Performance and Full Scale IQ scores were
stable over the three-year re-evaluation period. In order to determine whether certain groups would elicit less stable scores, several variables were investigated (special education classification, Full Scale IQ score at initial administration, age of the participants at initial administration, ethnicity and gender of the participants).

While investigating each variable, the groups were subdivided accordingly and three statistical analyses (correlational method, test of mean difference, and test of intra-individual variability) were employed to investigate stability. Presently, there are no published studies that have investigated the stability of WISC-III scores over the mandated three-year interval. Therefore, there is no means to generate hypotheses about the stability of WISC-III scores based on previous research. The hypotheses in this current study were generated by examining the trends that were prevalent during investigations of WISC and WISC-R stability studies. It is interesting to examine the comparability of results to see if similar trends exist for this new revision of the Wechsler scale.

**Stability of IQ Scores Based on Special Education Classification**

As a result of the initial administration of the WISC-III, the 592 children reported in this study were classified as learning disabled (n=467), not eligible for services (n=36), behavior disordered (n=28), speech/language disordered (n=24), mentally retarded (n=15), Attention-Deficit/Hyperactivity Disorder (n=10), and Otherwise Health Impaired (n=10).

**Learning disabled**

Children classified as learning disabled elicited high correlation coefficients for test-retest VIQ, PIQ, and FSIQ scores. No significant differences were found in this
investigation between mean test-retest VIQ, PIQ and FSIQ scores over the three year interval. Intra-individual variability was not significant for test-retest scores across the three IQ measures.

Based on previous research, it was predicted that the learning disabled sample would experience significant group mean increases in PIQ scores, and significant group mean decreases in VIQ scores between test-retest administrations. No significant differences were found between test-retest mean VIQ, PIQ nor FSIQ scores over the three-year interval. These results are inconsistent with previous research investigating the stability of the WISC and WISC-R. Significant mean increases in VIQ, and significant mean decreases in PIQ were reported for a group of learning disabled children administered the WISC twice over a fifteen month interval (Coleman, 1963). Similar trends were reported investigating the stability of the WISC-R over a three-year interval (Anderson, Cronin, & Kazmierski, 1988; Elliott & Boeve, 1987; Kaye & Baron, 1987; Haddad, Juliano & Vaughan, 1984). It was believed that increased Performance IQ scores in a learning disabled population was a result of practice effects (Covin, 1977). It was also suggested that decreased Verbal IQ scores was a result of the increased difficulty that handicapped children experience with verbal conceptualization and abstract verbal thinking as grade level advances (Naglieri & Pfeiffer, 1983). However, results of this study suggest that the WISC-III may be a more reliable instrument for the learning disabled population than the previous Wechsler scales. Test-retest VIQ, PIQ and FSIQ scores are stable for the learning disabled sample, and significant group mean differences are not expected as was the case with the WISC and WISC-R.
Mentally retarded

This investigation led to a very small sample of children classified as mentally retarded (n = 15). This sample demonstrated a non-significant correlation between test-retest Verbal IQ scores, a moderate correlation coefficient for test-retest Full Scale IQ scores, and a strong correlation coefficient for test-retest Performance IQ scores. Intra-individual variability was not significant across the three IQ measures for this population.

Based on previous research, it was predicted that the mentally retarded sample would not experience significant group mean differences in VIQ, PIQ and FSIQ scores between test-retest administrations. Results confirmed this hypothesis, whereby no significant mean differences were found between administrations for the sample of fifteen children classified as mentally retarded. These results are comparable to those reported investigating the stability of test-retest WISC scores (Throne, Schuman & Kaspar, 1962; Friedman, 1970) and test-retest WISC-R scores (Whorton, 1985) for this population.

Caution, however, should be taken when interpreting this study's results because of the very small sample size. To further investigate the stability of WISC-III IQ scores for the mentally retarded population, a larger sample must be utilized to interpret results with confidence.

It was hypothesized that the learning disabled sample would elicit less stable correlation coefficients for test-retest VIQ, PIQ and FSIQ scores than for the mentally retarded sample. Results indicated no significant difference between PPM correlation coefficients for test-retest PIQ and FSIQ between the two groups. However, the test-retest correlation coefficient for VIQ was more stable for the learning disabled sample than for the mentally retarded sample. This result contradicts the original hypothesis. The
unforeseen result may be explained due to the discrepant sample size. Four hundred and sixty-seven learning disabled children were compared to fifteen mentally retarded children. In order to adequately investigate this hypothesis, it is necessary to gather information on the two samples consisting of a comparable number of subjects.

**Behavior disordered**

Children classified as behavior disordered demonstrated high correlation coefficients across the three IQ measures. No significant differences were found between test-retest VIQ and FSIQ scores. However, a significant difference was found between test-retest PIQ scores. Scores decreased by more than four points between administrations. Such a decrease in scores may be a result of increased demands in completing tasks requiring significant attention and concentration. Such deficits in attention and concentration are common in children classified as behavior disordered. Intra-individual variability was not significant for test-retest VIQ and FSIQ. However, a mean difference score of 4 points for PIQ test-retest scores was significant. Because of the small sample reported in this study (n=28), caution should be taken in the interpretation of results. Future research should utilize a larger sample to further investigate the reliability of WISC-III scores for this population, specifically to explain the significant decreases in PIQ scores.

**Not eligible for services**

This group of thirty-six children was referred for an evaluation, administered the WISC-III and was considered not eligible for special education services. Approximately three years later, these children were again referred and re-administered the WISC-III. Children categorized as "not eligible" for services as a result of the initial administration had strong test-retest correlations for VIQ, PIQ and FSIQ scores. In addition, no
significant differences were found between test-retest scores across the three IQ measures. Intra-individual variability was also not significant between test-retest scores across the three IQ measures for this population. Future studies should be conducted to investigate this population of children who tend to be re-referred. Because this investigation utilized a small sample, further studies should confirm the findings reported with larger samples. If results are inconsistent, studies should investigate several factors that may influence the stability of scores for this population such as the child’s age at initial administration, performance on the initial administration, and regular education remedial services received, if any.

Attention-Deficit/Hyperactivity Disorder

Only ten children were classified as ADHD in this investigation. Correlation coefficients were moderate for test-retest VIQ, PIQ and FSIQ scores. No significant mean differences nor intra-individual variability were found between test-retest scores across the three IQ measures. Because correlation coefficients appear to be lower than other populations reported in this investigation, future studies with larger samples should be conducted to detect if children classified as ADHD tend to have less stable scores than children diagnosed with other special education classifications. In addition, it would be interesting to investigate several factors that may influence the reliability of WISC-III scores for this subgroup, such as the use of medication, the type of services received and the influence of double classifications (e.g. a child classified as ADHD and learning disabled).
Otherwise Health Impaired

Children classified as Otherwise Health Impaired also constituted a very small sample (n=10). High correlation coefficients for VIQ, PIQ and FSIQ scores were found. No significant differences nor intra-individual variability were discovered between test-retest mean VIQ, PIQ and FSIQ scores. Again, because of such a small sample, caution should be taken when interpreting these scores and further investigations should be conducted utilizing larger sample sizes. It would be interesting to investigate several factors that may influence the stability of IQ scores for this population, such as the type of services received and the specific impairment of the children under investigation.

Speech/language disordered

Twenty-four children were classified as speech/language disordered as a result of the initial WISC-III administration. Correlation coefficients were strong for test-retest VIQ, PIQ and FSIQ scores. No significant mean differences nor significant intra-individual variability were found between test-retest scores across the three IQ measures. Caution should be taken when interpreting these results because of the small sample investigated. Future studies should investigate this population with larger sample sizes. One factor that may influence the reliability of WISC-III scores which would be interesting to investigate is the relationship between speech/language services received, and the stability of VIQ scores over time.

Stability of IQ Scores Based on Initial Full Scale IQ Score

In order to investigate the effects of initial FSIQ score on the stability of test-retest scores, the sample was divided into three groups: children receiving an initial FSIQ below
90, children receiving a FSIQ between 90 and 109, and children receiving a FSIQ above 109.

Based on previous research investigating the stability of scores with the WISC and WISC-R, it was predicted that students receiving an initial FSIQ below 90 would have more stable correlation coefficients than subjects scoring above 109. Results confirmed these previous studies whereby children with initial FSIQ below 90 demonstrated significantly higher test-retest correlation coefficients for VIQ, PIQ and FSIQ scores than children scoring above 109 (Naglieri & Pfeiffer, 1983; Bauman, 1991; Klonoff, 1972).

It was also predicted that group mean differences for test-retest VIQ, PIQ and FSIQ scores would not be significant for students with initial FSIQ scores below 90. Results of this investigation disconfirmed this hypothesis. Results indicated that significant mean differences exist between test-retest administrations for this group. Children who had an initial FSIQ scores below 90 scored significantly higher on the second administration approximately three years later by a mean of 1.24, 1.65 and 1.54 on VIQ, PIQ and FSIQ, respectively. Such increases in scores are inconsistent with previous research of the WISC and WISC-R (Whatley & Plant, 1957; Naglieri & Pfeiffer, 1983).

It was also hypothesized that group mean differences for test-retest VIQ, PIQ and FSIQ would be significant for students with initial FSIQ scores above 109. Results of this study indicated that mean VIQ, PIQ and FSIQ scores significantly decreased over administrations by means of 3.12, 6.36 and 5.32 points, respectively, over the three-year period. Such results are consistent with previous research investigating the stability of the WISC-R (Bauman, 1991).
Children receiving an initial FSIQ below 90 tended to score significantly higher on the second administration, whereas children scoring above 109 tended to score significantly lower on the second administration three years later. Such results may be attributed to the phenomena of regression to the mean. However, the discrepancy in scores between administrations for the below 90 group appear to be pragmatically insignificant. An increase in scores by less than two points will not have a significant impact on the decision making process for a change in classification nor eligibility of services. However, the discrepancy in scores for the children who scored above 109 on FSIQ at the time of the initial administration has significant practical implications. Such decreases in scores by more than three, five and six points may play a significant role in the decision making process for a change in classification or eligibility of services.

Stability of IQ Scores Based on Age at Initial Administration

The present study compared differences and similarities between test-retest performance on the WISC-III at two age levels: below eight years of age, and eight years and older.

Based on previous research by Bauman (1991) with the WISC-R, it was predicted that test-retest correlation coefficients for VIQ, PIQ and FSIQ scores would be more stable for children eight years of age and older than for children under the age of eight at the time of the initial administration. Results of this current investigation contradicts previous research. No significant differences were found between the test-retest correlation coefficients for VIQ and FSIQ based upon the age of the participants at initial administration. Test-retest correlation coefficients for VIQ and FSIQ for both groups were strong. Differences between test-retest correlation coefficients were significant only
for PIQ scores, whereby children eight and older demonstrated more stable coefficients compared to children below age eight at initial administration. Such results may suggest that the WISC-III is a more reliable instrument for children under the age of eight for VIQ and FSIQ than previous Wechsler scales. Comparative analyses with the WISC, WISC-R and WISC-III should be conducted to confirm this finding. Future studies may also investigate the PIQ discrepancy in correlation coefficients between age groups. One investigation should examine the stability of the Coding subtest of the Performance Scale since this is the only subtest that is different for both groups.

Based on previous research with the WISC (Klonoff, 1972) and WISC-R (Bauman, 1991), it was predicted that group mean differences for test-retest VIQ, PIQ and FSIQ scores for children under eight years of age at initial administration would be significantly higher than the group mean differences for children eight years of age and older. Results indicated that a significant interaction effect exists suggesting that there is a significant differential change over time by age group. Further analysis of test-retest Verbal IQ scores suggested that children below age eight at initial administration improved by less than one point and children age eight and older decreased by less than one point between administrations. Results of the test-retest PIQ analysis suggested that children under age eight decreased by nearly two points, whereas children eight and older increased by less than one point. No significant differences were found between test-retest FSIQ scores between groups. Although these results are statistically significant, the influence of three years time on the VIQ and PIQ scores of this sample of handicapped children was pragmatically insignificant and may not be found useful for diagnostic purposes.
Results also suggested that for children under the age of eight at the time of the initial administration, significant intra-individual variability existed. PIQ scores increased by a mean of 1.80 points between administrations. No significant intra-individual variability existed for test-retest VIQ and FSIQ scores. For the children who were eight years of age and older at the time of the initial administration, significant intra-individual variability was only significant for VIQ scores whereby scores increased by a mean of .84 points. Such small increases and decreases in scores over the three year interval does not have practical implications. Therefore, age of participants at initial administration does not play a significant factor in the stability of WISC-III IQ scores over the three-year test-retest interval.

**Stability of IQ Scores Based on the Ethnicity of Participants**

Because of the small percentage of children categorized as non-Caucasian (17.2%), the analysis divided the entire sample into two groups, Caucasian and non-Caucasian. It was predicted that there would be no significant differences between the correlation coefficients from test-retest VIQ, PIQ and FSIQ scores based upon the ethnicity of the child. Previous research from Elliott and Boeve (1987) and Elliott et al. (1985) investigated the effects of ethnicity on the three-year WISC-R stability of the performance of children in special education. Results of both studies suggested that the variable ethnicity did not influence the test-retest VIQ, PIQ and FSIQ stability over time. A multivariate method was employed in the present study to investigate the effects of time and ethnicity on the stability of WISC-III scores for 492 Caucasian children, and 79 children categorized as non-Caucasian. Results confirmed previous research, whereby no significant differences were discovered between differential change scores based on the
ethnicity of the subjects. In addition, no significant differences between test-retest correlation coefficients were discovered based on the ethnicity of the subjects. Intra-individual variability was not significant across the three IQ measures for either subgroup. Thus, ethnicity is a variable that does not affect the stability of WISC-III scores over a three-year interval.

**Stability of IQ Scores Based on the Gender of Participants**

It was hypothesized that gender differences would not exist in the stability of WISC-III IQ scores over the three-year interval. Results of this investigation suggested that the only significant difference between correlation coefficients was that female test-retest PIQ scores were significantly more stable than test-retest PIQ scores for males over a three-year interval. Such results are surprising and inconsistent with previous research investigating WISC stability (Klonoff, 1972). Further investigation should examine whether true gender differences exist that may affect the stability of PIQ over time. However, one may also see such differences as minimal in that small changes in test-retest scores with very large sample sizes may explain these findings. Therefore, these results may not have practical implications. Utilizing a multivariate repeated measures analysis of variance, results suggested that no significant interaction exists between males and females in the amount of change between administrations for VIQ, PIQ and FSIQ scores. Intra-individual variability was not significant for test-retest scores across the three IQ measures for either subgroup under consideration. Therefore, VIQ, PIQ and FSIQ scores for the WISC-III are stable for males and females over the three-year test-retest re-evaluation period.
Conclusions

Results of this study conclude that for a homogeneous group of children in special education, WISC-III Verbal, Performance and Full Scale IQ scores remain stable over a three-year test-retest interval. However, test reliability appears to vary as a function of several factors.

Investigating the effects of special education classification on test-retest stability, results suggest that test-retest VIQ, PIQ and FSIQ scores are relatively stable for each group under consideration. The only significant test of mean difference and test of intra-individual variability found was that children classified as behavior disordered tend to score significantly lower on PIQ by four points between administrations. However, because each classification group was represented by such a small sample size (excluding the learning disabled group), further research should investigate each classification category under consideration with larger sample sizes to confidently determine whether classification plays a role in affecting the test-retest reliability of WISC-III scores over time.

In order to determine whether initial Full Scale IQ scores affect test-retest reliability, the sample was divided into below average, average and above average subgroups. Results of the current investigation suggest that children scoring a FSIQ below 90 at initial administration tend to have greater stability than children scoring above 109. Interestingly, it was also determined that children scoring below 90 tend to score significantly higher on the second administration, and children scoring above 109 tend to score significantly lower on the second administration. Such results can be attributed to the phenomena of regression to the mean. However, the increases in VIQ, PIQ and FSIQ
scores that were prevalent in the subgroup that scored below 90 on FSIQ at initial administration were very small. On average, increases in scores were less than two points across the three IQ measures. On the other hand, for the children that scored a FSIQ above 109 at initial administration, VIQ, PIQ and FSIQ scores decreased by more than three, six and five points, respectively. Therefore, children who score an initial FSIQ above 109 will tend to have less stable scores than children who score in the average or above average range.

Age does not play a significant factor in test stability of the WISC-III. Although results suggest that children eight and older at the time of the initial administration demonstrate more stable PIQ correlation coefficients than children under eight at the time of the initial administration, such differences are relatively small and not diagnostically useful. In addition, significant but minimal mean differences exist between test-retest VIQ and PIQ scores based upon the age of the participant. However, these results also appear pragmatically insignificant.

No significant differences were discovered between differential change scores based on the ethnicity of the participants. Similarly, no significant interaction exists between the gender of the participant and the amount of change between administrations. However, females demonstrated more stable PIQ scores than male participants. Such a significant but small difference in stability also appears to be pragmatically insignificant.

Limitations

Although this investigation is comprised of a very large sample of 592 children, several limitations are quite evident and affect the generalizability of the study. First, the sample of subjects was limited to a small New England state. This sample is not
representative of the United States population, and caution should be taken when generalizing these results. Second, because the criteria for classification differs among states, the results obtained may not be representative to children outside the state because their membership in the classification categories may differ. Third, small sample sizes were evident in examining the relationship of special education classification on the reliability of IQ scores. Larger samples are necessary in future research to interpret such results with confidence. Fourth, the ethnicity of the subjects do not appear to represent the national percentages of individuals in special education.

Implications

The current study has several implications for the practicing school psychologist with respect to the necessity of conducting cognitive assessments every three years. The present study concluded that WISC-III Verbal IQ, Performance IQ and Full Scale IQ scores are generally stable for children in special education over a three-year re-evaluation period. Therefore, the practice of conducting three-year re-evaluations utilizing the same cognitive assessment for every child is not viably useful.

This study investigated the stability of scores for several populations within the special education community to determine whether certain children will fluctuate in VIQ, PIQ and FSIQ scores over the re-evaluation period. Results of this investigation suggest that generally scores remain stable over time. Children classified as learning disabled, not eligible for services, ADHD, Otherwise Health Impaired or speech and language disordered will tend to have stable scores and re-evaluations do not appear necessary. However, children classified as mentally retarded tend to experience instability in VIQ scores over the three-year re-evaluation period. Similarly, children classified as behavior
disordered experience instability in PIQ scores over a three year interval. In addition, children who received an initial FSIQ score above 109 will tend to fluctuate significantly between administrations. Therefore, for these populations, re-evaluations appear justified.

Because significant differences are not expected for the rest of the population of children in special education, retesting the same child over three years will not yield new information. By gaining an accurate assessment of children's intellectual ability during the initial administration, further administrations of the same cognitive tool will contribute limited information to educational planning and program assessment for children in special education. Therefore, routine administrations of individually administered IQ tests, namely the WISC-III, is of questionable value and should no longer be required for all children.

The re-evaluation process consumes an extensive amount of school psychologists' time as well as financial resources of school districts. This study has concluded that routine re-evaluations are not necessary for all children and thus should be eliminated. The elimination of unnecessary re-evaluation assessments will result in financial relief of school districts who will no longer be responsible for financing the needless practice of conducting individual cognitive assessments. As a result, school districts will be able to disburse those funds more effectively. Such funding would aid in developing more appropriate programs for children in special education to service their educational and psychological needs. The elimination of routine re-evaluations will also have dramatic implications on the roles and functions of school psychologists. School psychologists will have more opportunity to perform the roles they are trained in such as counseling and teacher consultation as well as have a direct influence in the design and implementation of
educational and psychological services for children that are desperately needed in the schools today.

It is imperative that caution is taken when interpreting the results of this study. Policy makers are strictly warned not to carelessly make judgments regarding the necessity of school psychologists in the districts. The elimination of the re-evaluation process should not eliminate the employment of school psychologists in the schools. On the contrary, the release of such testing constraints will significantly expand the role of school psychologists. The elimination of testing will provide the opportunity for school psychologists to perform the jobs that they are intended.

It is inevitable that questions will be raised by school psychologists, school administrators and policy makers regarding the future role of school psychologists. Before decisions are made on a state or district level, it is imperative that school psychologists unite to develop a professional philosophy regarding the future of their profession. After such discussions, the State school psychology association must collaborate with school administrators and policy makers to discuss the practical implications of eliminating three-year re-evaluations and the effect such policies will have on the expanding role of the school psychologist.
APPENDIX A

Institutional Review Board
Letter of Approval
The University of Rhode Island

INSTITUTIONAL REVIEW BOARD ON HUMAN SUBJECTS (IRB)

IRB ACTION REPORT

The activity indicated below has been reviewed by the University of Rhode Island Institutional Review Board (IRB) in accordance with the requirements of Title 45, Part 46 of the Code of Federal Regulations (Protection of Human Subjects), or other federal regulations as required such as 21CFR 50. The University has an approved assurance of compliance on file with the Department of Health and Human Services which covers this activity. Our assurance number is M1457. Any changes which may alter the investigational situation must be reported promptly to the IRB. Any questions concerning this action can be directed to:

Barbara S. Brown
Director of Compliance
The Research Office
70 Lower College Road
University of Rhode Island
Kingston, RI 02881
telephone: (401) 874-4128

Date: August 6, 1996

<table>
<thead>
<tr>
<th>Project Title: “Stability of WISCIII Scores: For Whom are Triennial Re-evaluations Necessary?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty Investigator or Sponsor: Janet Kulberg</td>
</tr>
<tr>
<td>Psychology</td>
</tr>
<tr>
<td>Chafee Social Science Center</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Student Investigator of Investigation: Linda Cassidy</td>
</tr>
<tr>
<td>Psychology</td>
</tr>
<tr>
<td>Chafee Social Science Center</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Date of Initial IRB Review: May 20, 1996</td>
</tr>
<tr>
<td>Type of Review: Exempt</td>
</tr>
<tr>
<td>Date of Approval: 8/6/96</td>
</tr>
<tr>
<td>Action: Approved</td>
</tr>
<tr>
<td>Comment:</td>
</tr>
</tbody>
</table>

IRB Chair
(Or Designated Member)

Date

Barbara S. Brown
Director of Compliance

Date

81
APPENDIX B

Correspondence
June 14, 1996

TO: Directors of Special Education

FROM: Robert M. Psychod. Director
Office of Special Needs

RE: Triennial Evaluations Research Project

Ms. Linda Cassidy, with endorsement of the Rhode Island School Psychologists Association and the Department of Education, will be studying the stability of WISC III scores over time. As you know, the IDEA reauthorization may alter the requirements and format of triennial evaluations so Ms. Cassidy’s study is timely and should be informative for developing new policies.

She will be working with your school psychologists to obtain non-personally identifiable data which will not need any waiver of confidentiality requirements.

Please contact me at 277-2505x2201 if you have any questions.

RMP.dl
Dear Linda:

I am pleased to write this letter on behalf of RISPA. At our May 1996 RISPA Executive Board meeting we voted in support of your project "Stability of WISC III Scores: For Whom are Re-evaluations Necessary?". We understand that these data may be used to inform how IDEA reauthorization is enacted in Rhode Island, and the potential impact on our profession. In addition, we understand that your study is being supported by the Rhode Island Department of Education, Office of Special Needs, Robert Fryhoke, Director in the amount of $500.00.

We will provide you with the following in support of your project:

1. A letter of support to encourage participation of school psychologists in data collection in their school districts.
2. The RISPA mailing labels

Funding from the Office of Special Needs will be made payable to RISPA. In turn we will compensate participating school psychologists and participate-in school psychology students in the amount of $50 following your verification of individual completion of data collection.

Thank you for your commitment to Rhode Island children, their schools, and our profession. We look forward to the results of your study, and strongly urge you to present them at the RISPA Annual Summer Conference.

Sincerely,

Denise M. DeZolt, Ph.D.
President, RISPA
Dear Bob:

I am writing as a follow up to our meeting with Linda Cassidy on this date. Specifically, I will outline the RISPA support for Ms. Cassidy's research project, and detail my understanding of the agreement between RISPA and the Office of Special Needs.

At our May 1996 RISPA Executive Board meeting we voted in support of Ms. Cassidy's project: "Stability of WISC III Scores: For Whom are Re-evaluations Necessary?". We understand that these data may be used to inform how IDEA reauthorization is enacted in Rhode Island, and the potential impact on the school psychologist's role. In addition, I understand that this study is being supported by the Rhode Island Department of Education, Office of Special Needs in the amount of $5000.00.

RISPA will provide Ms. Cassidy with the following in support of her project:

1. A letter of support to encourage participation of school psychologists in data collection in their school districts.
2. The RISPA mailing labels

Funding from the Office of Special Needs will made payable to RISPA. In turn we will compensate in the amount of $50 participating school psychologists and practicum/intern school psychology students (working under direct supervision of district school psychologists and subject to the same confidentiality standards) following Ms. Cassidy's verification of individual completion of data collection.

It is a pleasure to collaborate with you on behalf of Rhode Island children, their schools, and our profession. To facilitate this effort, you may contact Judy Zeyl, RISPA Treasurer, at the address above. Following elections to the RISPA Executive Board, You will be notified of the contact person for continuing collaboration on IDEA reauthorization. Until then, I may be reached by phone or voice mail at 401-874-4221, or e-mail AGL101@URIACC.URI.EDU.

Sincerely,

Denise M. DeZolt, Ph.D.
President, RISPA
11 June 1996

Dear Colleague:

Here is an opportunity to earn $50 and be a part of shaping the role of school psychologists in Rhode Island. Please read on!!

At our May 1996 RISPA Executive Board meeting we voted in support of the project: "Stability of WISC III Scores: For Whom are Re-evaluations Necessary?" being conducted by Linca Cassidy, a doctoral student in school psychology at URI. Now we are encouraging you to participate by helping with data collection in your school district. These data may be used to inform how IDEA reauthorization is enacted in Rhode Island, and the potential impact on our profession. This study is being funded in part by the Rhode Island Department of Education, Office of Special Needs, Robert Fryhoda, Director. Mr. Fryhoda is communicating his endorsement to the Directors of Special Education throughout the state.

The enclosed letter from Ms. Cassidy details the data collection process. Should you choose to participate, you will receive $50 upon Ms. Cassidy's verification of your completion of the data collection:

Please consider including representation of your school district in this very important study. It is an opportunity to have a voice in the future of school psychology in Rhode Island.

If you have questions, you may contact Linca Cassidy at 401-762-3709, or a member of the RISPA Executive Board.

Sincerely,

[Signature]

Denise M. DeZolt, Ph.D.
President, RISPA
Dear Colleague,

I would like to personally invite you to participate in the study entitled, Stability of WISC-III Scores: For Whom are Triennial Re-evaluations Necessary? By assisting in the data collection process in your school district, you are helping to generate very important information that may ultimately assist school psychologists all over the state in determining which children would benefit from a psychological re-evaluation and which children might not need such an evaluation.

Endorsement and funding for this study has been provided by the Rhode Island Department of Education, Office of Special Needs. Bob Pryhoda, Director, is providing a letter to each of the Directors of Special Education in Rhode Island encouraging each district to participate in this study. The funding that has been provided has been allocated to you, the school psychologist, in assisting in the data collection process.

I understand the school year is coming to a close. However, I am asking each of you to take some time within the next couple weeks, or during the summer, to participate in this very important study (and make some money while you are at it) A sample data sheet is enclosed with this mailing. As you can see, we are seeking minimal information on the stability of WISC-III scores over time and therefore, collecting such information will not take you much time.

The only files that will be examined in this study are children who were administered the WISC-III twice. Because the WISC-III has only been published since 1991, you would only be examining the files of children who have been re-evaluated this past school year (1995-1996) and last year (1994-1995).

If you are interested in collecting this information and representing your district in this study whereby the results will be submitted to the Department of Education, please fill out the enclosed stamped postcard requesting all supporting material and instructions and return to me as soon as possible. If, however, you are unable to collect these data, I am asking your help in generating a list of children that were re-evaluated in your district in the past two years. In this manner, I will personally seek assistance into the files at the special education office this summer and conduct the file reviews personally. In either case, please return the enclosed postcard as soon as possible.

Thank you for taking the time to read this information. Any assistance you can give me at all would be greatly appreciated. If I can explain the research further to you or if you have any questions, please contact me at (401) 782-3709 or you can reach my major professor, Janet Kulberg, Ph.D., at (401) 874-4228. I look forward to hearing from you soon. Thank you.

Sincerely,

Linda Cassidy, M.S.
School Psychologist
<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________ 

<table>
<thead>
<tr>
<th>Student #</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Date</th>
<th>Age</th>
<th>Classification</th>
<th>Services received</th>
<th>Verbal IQ</th>
<th>Performance IQ</th>
<th>Full Scale IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments: ______________________


