1-1-2010

A Comparison Of The Wechsler Intelligence Scale For Children Third Edition And The Wechsler Intelligence Scale For Children Fourth Edition

Alyson M. Taylor
Eastern Illinois University

THESIS MAINTENANCE AND REPRODUCTION CERTIFICATE

TO: Graduate Degree Candidates (who have written formal theses)

SUBJECT: Permission to Reproduce Theses

The University Library is receiving a number of request from other institutions asking permission to reproduce dissertations for inclusion in their library holdings. Although no copyright laws are involved, we feel that professional courtesy demands that permission be obtained from the author before we allow these to be copied.

PLEASE SIGN ONE OF THE FOLLOWING STATEMENTS:

Booth Library of Eastern Illinois University has my permission to lend my thesis to a reputable college or university for the purpose of copying it for inclusion in that institution's library or research holdings.

Allyson M. Taylor _____ 05/04/2010 _____

Author's Signature Date

I respectfully request Booth Library of Eastern Illinois University **NOT** allow my thesis to be reproduced because:

_____

_____

_____

_____ _____

Author's Signature Date

This form must be submitted in duplicate.

A Comparison of the Wechsler Intelligence Scale for Children Third Edition and the Wechsler Intelligence Scale for Children Fourth Edition

BY

Alyson M. Taylor

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Specialist in School Psychology

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

2010
YEAR

I HEREBY RECOMMEND THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED ABOVE

THESIS COMMITTEE CHAIR — DATE 5/4/10	DEPARTMENT/SCHOOL CHAIR OR CHAIR'S DESIGNEE — DATE 5/4/10
THESIS COMMITTEE MEMBER — DATE 5/4/10	

Table of Contents

Acknowledgements

The completion of this project signifies the culmination of three years of time and dedication to the study of school psychology. There have been several people whose support and encouragement have been much appreciated and vital to my success. First, I'd like to thank my thesis chair, Dr. Gary L. Canivez, for his guidance throughout this process and for always pushing me to do my best. I would also like to thank my other committee members, Dr. Mike Havey and Dr. Ronan Bernas, for their support and feedback on this project. The entire faculty and the members of my cohort have been a huge factor in my success; they have been like my family throughout these past three years, always encouraging me. Last, even though my family did not have a direct influence on this project, they were always there with words of wisdom and comfort when I couldn't see the light at the end of the tunnel. To everyone who has played a part to get me to where I am today, a sincere thank you.

List of Figures and Tables

Abstract

There are no specific guidelines to determine when or why intelligence tests should be revised; however, revisions typically occur every 10 to 15 years to keep up with the moderate change in IQ over time, known as the Flynn effect (Flynn, 1984, 1987). Revisions also eliminate and update outdated items, content, and materials; extend age ranges, floors, and ceilings; as well as incorporate new theoretical changes. To assess cognitive abilities the Wechsler Scales are frequently used by clinical and school psychologists. Previous studies comparing versions of the Wechsler Intelligence Scale for Children (WISC) found significant correlations (.80s and .90s) and significant mean differences with the older version having higher scores than the newer version, providing evidence for the Flynn effect. The purpose of the current study was to examine the comparability of the WISC-III (Wechsler, 1991) and the WISC-IV (Wechsler, 2003) as well as find evidence of the Flynn effect. Participants were 89 students with Individualized Education Plans who underwent special education reevaluations in which the WISC-III was used at Time 1 and the WISC-IV at Time 2. Scores were examined using Pearson product moment correlation analysis, and dependent *t*-tests for differences between means were conducted to examine the significance between scores. Results found statistically significant correlations between similar or identical scores except for the verbal-performance difference scores. Only the mean differences for the WISC-III and WISC-IV FSIQ and PSI were significant, providing evidence for the Flynn effect. Future research should look to replicate this study with a larger and more representative sample of the general population.

A Comparison of the Wechsler Intelligence Scale for Children Third Edition and the Wechsler Intelligence Scale for Children Fourth Edition

The problem investigated in the current study was the comparability of the Wechsler Intelligence Scale for Children Third Edition (WISC-III; Wechsler, 1991) and the Wechsler Intelligence Scale for Children Fourth Edition (WISC-IV; Wechsler, 2003) in a sample of students referred for special education reevaluations. Comparability was specifically examined for students who underwent two consecutive special education evaluations in which the WISC-III was used at Time 1 and the WISC-IV was used at Time 2. These situations are especially important because the accuracy of the results is brought into question when determining special education eligibility. After the publication of the WISC-IV, school psychologists had to compare previous results on the WISC-III with results on the WISC-IV.

There are no set ethical, practical, or legal guidelines in regard to when or why a test should be revised. In the past, tests underwent more revisions in a shorter amount of time (American Psychological Association [APA], 1966). However, due to the work of Flynn (1984; 1987), intelligence tests are typically updated every 10 to 15 years in order to be useful for one generation of individuals (Adams, 2000). Tests should be revised when a newer version will give a better indication of the construct being measured and when a newer version will better differentiate between individual levels of performance (Silverstein & Nelson, 2000). According to *The Standards for Educational and Psychological Testing*:

> Revisions or amendments are necessary when new research data, significant changes in the domain, or new conditions of test use and interpretation would

> either improve the validity of interpretation of the test scores or suggest that the test is no longer fully appropriate for its intended use. (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999, p. 42).

Tests are typically revised to update the normative sample and eliminate and update outdated items, content, and materials, as well as extend age ranges, floors, and ceilings (Adams, 2000; Strauss, Spreen, & Hunter, 2000). Changes in content are particularly necessary in order to ensure test security (AERA, APA, & NCME). Tests are also revised to improve the psychometric properties such as reliability and validity. Some tests are revised to address the basic assumptions and theories that guided the development of the test in the first place (Adams). It is important to understand how test revisions compare with the previous version of the test because neglecting to do so may lead to improper results and interpretations, such as inflated IQ scores on the older version (Strauss et al.).

Intelligence tests are also revised and renormed to keep up with the moderate change in IQ over time, known as the Flynn effect (Flynn, 1984, 1987). The Flynn effect is a phenomenon first discovered in 1984 by James Flynn that describes a progressive increase in IQ over time, detected by significant mean differences between scores on IQ tests. Significant mean differences are found when a test has outdated norms (after approximately 10 years) or when a test has undergone a revision. Studies have shown that the Flynn effect is real and meaningful. Steen (2009) performed a meta-analysis of 16 studies, all of which confirmed that the Flynn effect exists in 12 countries (1 developing 11 developed) in divergent cultures using a wide variety of tests. Findings also extended across a wide age range of individuals. To date no studies have found a specific cause of

the Flynn effect. Hypotheses suggest it is likely due to a combination of nature and nurture: the brain is constantly evolving, child development is happening sooner than in the past, children are healthier nowadays, and family environments are more supportive of children and better able to foster cognitive growth (Steen). Flynn believes a good explanation for the increase in IQ is due to the switching from more concrete to more abstract and scientific thinking (Shalizi, 2009).

Although there is no definite answer to the question of how large the Flynn effect is, most studies find that the average gain is approximately three points per decade (Steen, 2009). As a result, intelligence tests are renormed to adjust for these gains and to keep the average score for an age group at 100. The Flynn effect causes problems when interpreting scores on the outdated test, as well as comparing performance from one version of a test to another. As a test ages, children tend to obtain scores that are not indicative of their true intellectual functioning in comparison to peers their age because the normative sample no longer adequately represents the children. Due to the Flynn effect, the periodic updating of norms for tests assessing cognitive abilities is necessary because average IQ scores in the population gradually drift upward and give an increasingly deceptive picture of an individual's abilities in comparison to others of the same age. This renorming causes significant mean differences between scores on the newer and older version of an intelligence test. On average, there is approximately a three to six point score difference, with scores on the revision being lower in value (Flynn, 1984, 1987).

Even taking into account the Flynn effect, intelligence is a psychological construct considered to be stable over time; therefore, any test designed to measure

intelligence must produce similar scores from one testing to another. This is especially important as it is commonplace for school psychologists to readminister comprehensive intellectual tests as part of triennial reevaluations for special education (Moffitt, Caspi, Harkenss, & Silva, 1993). However, between evaluations, it is possible that the intelligence test used in the first evaluation underwent a revision.

When an intelligence test is revised, it is important that the new version and the previous edition lead to similar scores. If similar scores are not obtained, it could be due to types of test items, placement of test items, and administration and scoring techniques. Basically, any difference between versions can result in different scores. Regression toward the mean is yet another reason why scores may differ as test scores are not perfectly reliable (Sattler, 2001). Scores that were initially higher tend to be lower on the second testing and scores that were low on the first testing tend to be higher on the second testing. Scores drift toward the average. Some suggest it is important to determine whether the new version of the test is more sensitive at recognizing dysfunction than its predecessor; is the new test better at distinguishing between a typical child and a child with a learning disability or a child with cognitive impairment (Strauss et al., 2000)? Score differences between tests can arise due to softening norms, which are those that are no longer adequately representative of the population.

To assess cognitive abilities, the Wechsler Scales are frequently used by both clinical psychologists and school psychologists. Goh, Teslow, and Fuller (1981) found within the area of intelligence testing, the three Wechsler Scales were in the top eight most frequently used tests. The Wechsler Intelligence Scale for Children-Revised (WISC-R) was the test most often used. The Wechsler Adult Intelligence Scale (WAIS)

ranked fourth, and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) ranked seventh (Goh et al.). Hutton, Dubes, and Muir (1992) conducted a comparable study and, again, found that the WISC-R, the WAIS, and the WPPSI were three of the most frequently used assessment instruments by school psychologists. Stinnett, Havey, and Oehler-Stinnett (1994) surveyed a random sample of members of the National Association of School Psychologists regarding their current assessment practices. Results demonstrated that school psychologists spent approximately half of their time conducting assessments and the Wechsler Scales were among the most frequently used assessment instruments. Watkins, Campbell, Nieberding, and Hallmark (1995) conducted a similar study with clinical psychologists, and of the instruments used most often, the WAIS-R and the WISC-III ranked second and third, respectively.

WB-I—WISC

The Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) was published as a downward extension of the Wechsler-Bellevue Form I (WB-I) test of cognitive abilities for adults. Since the WB-I was used to assess to cognitive abilities of adults and older children, the WISC was developed in order to have an intelligence test for younger children. The WISC adapted the Information, Arithmetic, Similarities, Vocabulary, Digit Span, Comprehension, Picture Comprehension, Picture Arrangement, Block Design, Object Assembly, and Coding subtests from the WB-I to use with children. To complete the WISC, another subtest, Mazes, was developed. The WISC was divided into Verbal and Performance Scales. It was possible to derive three different IQ scores from the WISC: a Verbal IQ (VIQ), a Performance IQ (PIQ), and a Full Scale IQ (FSIQ) (Wechsler, 2003). Since both the WB-I and the WISC were assumed to be

equivalent measures of IQ for the overlapping population of individuals aged 10 to 15, it was important to examine this assumption. Delattre and Cole (1952) examined the relationship between the WB-I and the WISC. Subjects in the study were 50 children attending public schools near Occidental College. They ranged in age from 10 years 5 months (10-5) to 15-7, and over half were female. Results indicated a correlation of .86 for the VIQ, .82 for the PIQ, and .87 for the FSIQ. Mean differences between tests were not reported. Price and Thorne (1955) found similar correlations between the WB-I and the WISC with a sample of 40 White American children at two age levels containing equal numbers of boys and girls. Both tests were administered to the students in the same test session with 15 minutes between tests. For children with a mean age of 11- 6, correlations were .85, .79, and .89 for the VIQ, PIQ, and FSIQ, respectively. Correlations for children with a mean age 14-6 were .90 for the VIQ, .41 for the PIQ, and .78 for the FSIQ. At both age groups, higher mean IQs were obtained on the WISC, which is a finding opposite of the Flynn effect. Mean differences between the WISC and the WB-I were not tested for significance. Due to the short amount of time between administrations, practice effects were particularly notable on the PIQ. See Table 1 for a summary of the studies comparing the WB-I to the WISC.

Table 1

Summary of WB-I—WISC Studies

Study	Sample	Order	Correlations	Mean Differences
Delattre and Cole (1952)	50 children in public schools Age 10 to 15	WISC given first to 48 students	.86 for the VIQ .82 for the PIQ .87 for the FSIQ	Not reported

Price and Thorne (1955)	40 White American children	Counterbalanced 15 min between tests	Age 11-6 .85 for the VIQ .79 for the PIQ .89 for the FSIQ Age 14-6 .90 for the VIQ .41 for the PIQ .78 for the FSIQ	Not reported

WISC—WISC-R

The Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) kept the original 12 subtests from the WISC, but shifted the age range from 5 to 15 to 6 to 16 (Wechsler, 2003). There were a great number of improvements over the WISC. Several subtests were lengthened in order to enhance reliability. Floors were lowered and ceilings were raised. Items believed to be culturally biased or out-of-date were revised and nonwhite children were added to the normative sample. Additionally, alternate administration of verbal and performance subtests became a part of the standardized administration. The comparability of WISC and WISC-R scores was not discussed in the WISC-R manual.

Schwarting (1976) compared the WISC and the WISC-R with a sample of 58 children who were randomly selected from a suburban school in Omaha, NE. Participants ranged in age from 6 years to 15 years at both testings, but mean ages were not reported. Tests were administered in a counterbalanced order, meaning some participants were given the WISC first and some were given the WISC-R first. Time between testings ranged from 60 to 67 days. Results indicated significantly higher IQ scores on the WISC ($p < .01$). Mean VIQ scores were 109.48 ($SD = 13.65$) for the WISC and 104.62 ($SD =$

14.08) for the WISC-R. Mean PIQ scores were 115.14 (*SD* = 14.56) on the WISC and 106.40 (*SD* = 15.18) on the WISC-R. Mean FSIQ scores were 113.40 (*SD* = 13.60) for the WISC and 105.91 (*SD* = 14.35) for the WISC-R. The differences between scores were reflective of the Flynn effect. VIQ, PIQ, and FSIQ differences all represented large effect sizes. Correlations between the WISC and the WISC-R were not calculated.

Hamm et al. (1976) compared the WISC and the WISC-R among a sample of 48 students who were from Educable Mentally Retarded (EMR) classes in Georgia. Subjects were divided into two groups on the basis of age. Students in Group I ranged in age from 9-6 to 10-6, and students in Group II ranged in age from 12-6 to 13-6. The average interval between testings was 39 days. The WISC-R was administered first to 34 children, and the WISC first to the remaining 14 children. Results indicated VIQ, PIQ, and FSIQ scores were significantly lower on the WISC-R than on the WISC for both age groups ($p < .001$). Mean VIQ scores for Group I were 71.77 (*SD* = 8.5) for the WISC and 64.95 (*SD* = 9.3) for the WISC-R. Mean PIQ scores for Group I were 76.77 (*SD* = 15.2) on the WISC and 67.32 (*SD* = 14.8) on the WISC-R. Mean FSIQ scores for Group I were 71.59 (*SD* = 11.1) for the WISC and 63.59 (*SD* = 12.1) for the WISC-R. Mean VIQ scores for Group II were 69.50 (*SD* = 7.2) for the WISC and 64.19 (*SD* = 7.7) for the WISC-R. Mean PIQ scores for Group II were 74.96 (*SD* = 13.2) on the WISC and 65.69 (*SD* = 12.0) on the WISC-R. Mean FSIQ scores for Group II were 69.42 (*SD* = 9.8) for the WISC and 62.23 (*SD* = 10.0) for the WISC-R. WISC—WISC-R correlations were .86, .83, and .89 for the VIQ, PIQ, and FSIQ, respectively.

Solly (1977) compared WISC and WISC-R scores for mentally retarded children and gifted children. There were 12 students in each group, and in each group there were 6

males and 6 females. Children ranged in age from 8-2 to 12-5. Participants were administered the WISC and the WISC-R 72 hours apart, using a counterbalanced design. Results were consistent with the Flynn effect, showing significantly higher mean WISC FSIQ scores for both groups ($p < .001$). Mean FSIQ scores for the gifted students were 136.08 on the WISC and 123.67 on the WISC-R. Mean FSIQ scores for the mental retardation sample were 76.25 on the WISC and 65.42 on the WISC-R. Correlations between the WISC and the WISC-R were not reported.

Swerdlik (1978) compared the scores on the WISC and the WISC-R in a sample of 164 Black, White, and Latino children who had been referred to the school psychologist due to concerns about their cognitive functioning. Tests were administered in a counterbalanced order with a test-retest interval between one week and one month. Results demonstrated that VIQ, PIQ, and FSIQ scores were significantly lower on the WISC-R ($p < .0001$). On the WISC, the mean VIQ was 86.83, the mean PIQ was 95.84, and the mean FSIQ was 91.33. On the WISC-R, the mean VIQ was 81.86, the mean PIQ was 89.96, and the mean FSIQ was 85.86. Overall mean WISC—WISC-R differences were similar for all ages ($p < .01$). WISC—WISC-R correlations were .90, .87, and .92 for the VIQ, PIQ, and FSIQ, respectively.

McGinley (1981) examined the relationship between scores on the WISC and the WISC-R for children who were "mentally handicapped." Participants ranged in age from 8 to 15-6. The interval between testings was between 7 and 28 days, and the two tests were administered in a counterbalanced manner. The mean WISC FSIQ was 68.05, and the mean WISC-R FSIQ was 61.57, and the difference between means was statistically significant ($p < .001$).

While some studies administered the WISC and the WISC-R in a counterbalanced manner, other studies examined reevaluation cases where the WISC was administered at Time 1 and the WISC-R was administered at Time 2. A revelation format, although used in practice, introduces error variance of the scores due to a time delay. Gironda (1977) used a reevaluation format and compared the WISC and WISC-R results with a sample of students attending EMR classes in New Jersey. Participants were mainly Black and there were twice as many boys as girls. At the time of the WISC administration the mean age was 11-0 and at the time of the WISC-R administration the mean age was 14-0. The retest interval ranged from six months to six years, with a mean interval of three years. No significant mean differences were found between the WISC and the WISC-R on the VIQ, PIQ, or FSIQ. The mean WISC-R VIQ score was 65.2 (SD = 7.58) and the mean WISC VIQ score was 67.2 (SD = 5.73). On the WISC-R, the mean PIQ score was 67.1 (SD = 13.65) and the mean WISC PIQ was 66.7 (SD = 7.98). The mean WISC-R FSIQ was 64.0 (SD = 9.52) and the mean WISC FSIQ was 63.9 (SD = 5.27). The WISC and the WISC-R VIQs were not significantly correlated (r = .21), but the PIQ correlation (r = .65) and the FSIQ correlation (r = .54) were statistically significant. The shared variance between VIQs was 4%, between PIQs was 42%, and between FSIQs was 29%.

Reschly and Davis (1977) also examined the comparability of the WISC and the WISC-R among students from the borderline and educable (mild) levels of mental retardation. Participants were between the ages of 7-9 and 16-1. Time between testings ranged from 5 months to 26 months. Findings indicated lower scores on the WISC-R for almost all comparisons. The WISC VIQ (M = 77.28, SD = 10.74) was higher than the WISC-R VIQ (M = 70.38, SD = 10.98). The difference between the WISC PIQ (M =

80.10, *SD* = 11.78) was not statistically different than the WISC-R PIQ (*M* = 80.06, *SD* = 14.26). The WISC FSIQ (*M* = 76.65, *SD* = 10.61) was higher than the WISC-R FSIQ (*M* = 73.04, *SD* = 12.20). The largest discrepancy was between scores on the VIQ. When comparing VIQ scores with PIQ scores, results were consistent with the general trend that PIQ scores were higher than VIQ scores. The correlations between the WISC and the WISC-R VIQ, PIQ, and FSIQ were .83, .80, and .87, respectively. Another study comparing scores on the WISC and WISC-R with a sample of students with mental retardation was conducted by Spitz (1983) with 33 students. The mean age for testing with the WISC was 12.92 and the mean age for testing with the WISC-R was 15.22. The mean interval between testings was 2.30 years. The mean WISC FSIQ was 61.42, while the mean WISC-R FSIQ was 56.20 ($p < .001$). The WISC—WISC-R FSIQ correlation was .70 ($p < .001$).

McGonagle (1977) examined the relationship between WISC—WISC-R scores in a referred clinical sample. Participants were 58 students in a suburban school district who were referred for psychological evaluation. Students were classified as EMR, Learning Disabled (LD), or they were receiving services in the regular education setting. Students were initially tested with the WISC and then reevaluated with the WISC-R due to state and federal requirements for students in special education or as a second referral for students in regular education. The age range when tested with the WISC was 6 years to 14 years, and the age range for the WISC-R was 8 years to 16 years. The time between testings ranged from two years to six years (*M* = 3.7). Results demonstrated significant correlations between all three scales, ranging from .56 for the PIQ for EMR students to .83 for the VIQ for students being served in the regular education class. In examining the

total sample, WISC-R IQs were significantly lower than the WISC IQs. One exception was EMR students' PIQ scores, which did not significantly differ.

Weiner and Kaufman (1979) examined the WISC versus the WISC-R in children with learning or behavioral disorders. Participants were 46 Black children, mainly boys between the ages of 7 and 10 who were referred to a Brooklyn, NY clinic. Correlations between the WISC and the WISC-R VIQ, PIQ, and FSIQ were .90, .82, and .90, respectively. A comparison of the means revealed that IQ scores on the WISC-R were significantly lower (around 7 points for the VIQ and around 8 points for the PIQ and FSIQ). Refer to Table 2 for a summary of WISC—WISC-R studies.

Table 2

Summary of WISC—WISC-R Studies

Study	Sample	Order	Correlations	Mean Differences
Schwarting (1976)	58 students from a suburban school in Omaha, NE Participants ranged in age from 6 years to 15 years at both testings	Counterbalanced 60 to 67 days between testings	Not calculated	Results indicated significantly higher IQ scores on the WISC ($p <$ .01) All IQ scores within the average range (85-115) VIQ, PIQ, and FSIQ differences all represented large effect sizes

Hamm et al. (1976)	48 students from Educable Mentally Retarded (EMR) classes in Georgia Students in Group I ranged in age from 9-6 to 10-6 Students in Group II ranged in age from 12-6 to 13-6	Counterbalanced Mean interval of 39 days between testings	.86 for the VIQ .83 for the PIQ .89 for the FSIQ	VIQ, PIQ, and FSIQ scores were significantly lower on the WISC-R than on the WISC for both age groups ($p <$ .001)
Solly (1977)	Mentally retarded children and gifted children Age 8-2 to 12-5	Counterbalanced 72 hours between testings	Not reported	Significantly higher mean WISC FSIQ scores for both groups ($p < .001$)
Swerdlik (1978)	164 Black, White, and Latino children referred due to concerns about their cognitive functioning	Counterbalanced	.90 for the VIQ .87 for the PIQ .92 for the FSIQ	VIQ, PIQ, and FSIQ scores were significantly lower on the WISC-R ($p <$ .0001)
McGinley (1981)	"mentally handicapped" Participants ranged in age from 8 to 15-6.	Counterbalanced	Not reported	The difference between means was statistically significant ($p <$ .001) The WISC-R FSIQ was lower by 7 points

Gironda (1977)	Mostly black students attending EMR classes in New Jersey Time 1 age 11 Time 2 age 14	Reevaluation	.21 for the VIQ .65 for the PIQ .54 for the FSIQ	No significant mean differences were found
Reschly and Davis (1977)	Borderline and educable (mild) levels of mental retardation Participants were between the ages of 7-9 and 16-1	Reevaluation	.83 for VIQ .80 for PIQ .87 for FSIQ	Lower scores on the WISC-R for almost all comparisons (not PIQ)
Spitz (1983)	Students with mental retardation Mean age for WISC was 12.92 and mean age for WISC-R was 15.22	Reevaluation	.70 for FSIQ	The WISC FSIQ was higher by 5 points. The mean WISC FSIQ was 61.42, while the mean WISC-R FSIQ was 56.20 ($p < .001$)
McGonagle (1977)	58 students in a suburban school district Students were classified as EMR, Learning Disabled (LD), or they were receiving services in the regular education setting Ages 6 to 16	Reevaluation Mean interval of 3.7 years	Significant correlations between all three scales, ranging from .56 for the PIQ for EMR students to .83 for the VIQ for students being served in the regular education class	In examining the total sample, WISC-R IQs were significantly lower than the WISC IQs. One exception was EMR students' PIQ scores, which did not significantly differ

Weiner and Kaufman (1979)	Black children from Brooklyn NY with learning or behavioral disorders	Reevaluation	.90 for the VIQ .82 for the PIQ .90 for the FSIQ	IQ scores on the WISC-R were significantly lower (around 7 points for the VIQ and around 8 points for the PIQ and FSIQ)

WISC-R—WISC-III

The third edition of the WISC (WISC-III; Wechsler, 1991) retained all 12 subtests from the WISC-R, and introduced an additional subtest, Symbol Search. In addition to the VIQ, PIQ, and FSIQ, four factor based scores were introduced in order to represent more factorially pure measures of cognitive functioning and were equivalent to the Cattell-Horn-Carroll (CHC; McGrew, 2005) "broad" dimensions of intelligence. These four factor scores included: the Verbal Comprehension Index (VCI), the Perceptual Organization Index (POI), the Freedom from Distractibility Index (FDI), and the Processing Speed Index (PSI). In addition to the typical reasons for updating a test such as updating norms and strengthening psychometric properties, the WISC-III was developed in order to revise and add items and subtests, as well as give the test a more contemporary appearance (Sattler, 2001). As part of the standardization, the WISC-R was compared to the WISC-III, with tests administered using a counterbalanced design to 206 children between the ages of 6 and 16. The sample consisted of 55% girls and 45% boys, with the majority of participants being White. The test-retest interval ranged from 12 to 70 days (median = 21). The results demonstrated significantly lower scores on the WISC-III. The mean WISC-III VIQ was 101.5 (*SD* = 14.5), and the mean WISC-R VIQ was 103.9 (*SD* = 14.7). The mean WISC-III PIQ was 104.2 (*SD* = 15.1), and the mean WISC-

R PIQ was 111.6 (*SD* = 15.4). The mean WISC-III FSIQ was 102.9 (*SD* = 14.7), and the mean WISC-R FSIQ was 108.2 (*SD* = 15.1). Correlations between the tests were .90, .81, and .89 for the VIQ, PIQ, and FSIQ, respectively (Wechsler, 1991).

Sevier and Bain (1994) examined the comparability of the WISC-R and the WISC-III among gifted students. The sample included 35 students in grades two through six. The retest interval was approximately 12 months. VIQ, PIQ, and FSIQ scores were significantly higher on the WISC-R than the WISC-III. The greatest difference was on the VIQ, with a mean difference of 14.57 points. Score differences in this study were higher than those reported in the WISC-III manual. Correlations between the two versions were lower, and were .57 for the VIQ, .34 for the PIQ, and .45 for the FSIQ.

Bolen, Aichinger, Hall, and Webster (1995) reviewed archival special education files of 136 students. Of the 136 files, 61 students had recently been administered the WISC-III as part of a special education reevaluation for continued eligibility and were included in the study. Time between testings ranged from two and a half years to three years. Approximately half of the subjects were White and half were Black. Students were classified with various disabilities including learning disability, educable mentally handicapped, behavioral emotional handicapped, and "at-risk." Results demonstrated significantly lower scores on the WISC-III ($p < .001$). Mean VIQs were 81.02 (*SD* = 14.65) for the WISC-R and 75.82 (*SD* = 14.92) for the WISC-III. Mean PIQs were 89.30 (*SD* = 15.55) on the WISC-R and 80.09 (*SD* = 16.01) on the WISC-III. Mean FSIQs were 83.74 (*SD* = 14.86) for the WISC-R and 75.79 (*SD* = 15.62) for the WISC-III. The VIQ-PIQ difference on the WISC-III was also smaller than the VIQ-PIQ difference on the WISC-R. Mean WISC-R—WISC-III differences were 5.20, 9.21, and 7.95 for the VIQ,

PIQ, and FSIQ, respectively. Correlations of the IQs on the WISC-R and the WISC-III ranged from .84 to .88, individual correlations were not listed.

Slate and Saarnio (1995) compared WISC-R and WISC-III scores for 257 children who were reevaluated for special education. All students were Caucasian. Participants were classified with Specific Learning Disability (SLD), Mental Retardation (MR), or another disability. The WISC-R—WISC-III correlations for the entire sample were .81 for the VIQ, .80 for the PIQ, and .84 for the FSIQ. Mean IQ differences on the FSIQ were 7.2 points, on the VIQ 5.8 points, and on the PIQ 7.5 points. Individual changes in IQs ranged from a decrease of 12 points to an increase of 29 points. Results indicated that the changes in IQ scores were greater for the MR students than for the rest of the sample, which was consistent with the results presented in the WISC-III manual (Wechsler, 1991).

Lyon (1995) compared the WISC-R and the WISC-III in learning disability reevaluations of 40 students, ranging in age from 6-9 to 12-7 at first testing. Children were primarily White and lower-middle to upper-middle class. Results found statistically significant lower WISC-III VIQ, PIQ, and FSIQ scores ($p < .001$). Correlations between the WISC-R and the WISC-III were .76, .56, and .85 for the VIQ, PIQ, and FSIQ, respectively.

Hamm et al. (1996) also compared the WISC-R and the WISC-III for special education reevaluations. Participants were 166 students between the ages of 6 and 16 who had been assessed with the WISC-R and three years later reassessed with the WISC-III. All students were receiving special education services in a resource setting. The average age at first testing was 9.8. Equal numbers of students were from rural and urban areas.

WISC-R—WISC-III correlations were .91, .89, and .94 for the VIQ, PIQ, and FSIQ, respectively. Results indicated consistently lower scores on the WISC-III than on the WISC-R. There was a mean difference of 1.68 points on the VIQ ($p < .02$), 6.47 points on the PIQ ($p < .0001$), and 4.57 points on the FSIQ ($p < .0001$). Coefficients of determination (r^2) were calculated in order show the amount of variance shared by the two tests. For the VIQ, the two tests shared 83% of the variance; for the PIQ, the two tests shared 74% of the variance; and for the FSIQ, the two tests shared 88% of the variance. See Table 3 for a summary of WISC-R—WISC-III studies.

Table 3

Summary of WISC-R—WISC-III Studies

Study	Sample	Order	Correlations	Mean Differences
Wechsler (1991)	206 children between the ages of 6 and 16	Counterbalanced Median of 21 days between testings	.90 for the VIQ .81 for the PIQ .89 for the FSIQ	Mean WISC-III IQs were lower
Sevier and Bain (1994)	Gifted students	Reevaluation 12 months between testings	.57 for the VIQ .34 for the PIQ .45 for the FSIQ	VIQ, PIQ, and FSIQ scores were significantly higher on the WISC-R than the WISC-III

Bolen, Aichinger, Hall, and Webster (1995)	Students labeled learning disability, educable mentally handicap, behavioral emotional handicapped, and "at-risk"	Reevaluation (examined special education files)	Correlations ranged from .84 to .88, individual correlations were not listed	Significantly lower scores on the WISC-III ($p <$ .001) Mean WISC-III—WISC-R differences were 5.20, 9.21, and 7.95 for the VIQ, PIQ, and FSIQ, respectively VIQ-PIQ difference on the WISC-III was smaller than VIQ-PIQ difference on the WISC-R
Slate and Saarnio (1995)	257 Caucasian children who were reevaluated for special education	Reevaluation	.81 for the VIQ .80 for the PIQ .84 for the FSIQ	Mean differences on the FSIQ were 7.2 points, on the VIQ 5.8 points, and on the PIQ 7.5 points
Lyon (1995)	Students with a learning disability	Reevaluation	.76 for the VIQ .56 for the PIQ .85 for the FSIQ	Statistically significant lower WISC-III VIQ, PIQ, and FSIQ scores ($p <$ .001)
Hamm et al. (1996)	All students were receiving special education services in a resource setting	Reevaluation	.91 for the VIQ .89 for the PIQ .94 for the FSIQ	Lower scores on the WISC-III than on the WISC-R Mean difference of 1.68 points on the VIQ ($p <$ .02), 6.47 points on the PIQ ($p <$ .0001), and 4.57 points on the FSIQ ($p <$ .0001)

WISC-III—WISC-IV

The Wechsler Intelligence Scale for Children Fourth Edition (WISC-IV; Wechsler, 2003) included greater revision than other versions of the Wechsler Scales in order to update the instrument's theoretical foundations, enhance clinical utility, increase developmental appropriateness, improve psychometric properties, and increase user-friendliness. The WISC-IV still provides a measure of general intellectual functioning (FSIQ), but the dual IQ model (VIQ and PIQ) from the previous versions is no longer utilized. Although the four factor model still exists, names of two of the indexes were changed. The POI on the WISC-III was changed the Perceptual Reasoning Index (PRI) to reflect enhanced emphasis on fluid reasoning abilities. The FDI was changed to the Working Memory Index (WMI) which gave a more accurate depiction of the abilities purportedly measured by this factor. The WISC-IV includes 15 subtests, which is more than any of the previous versions of this test. The subtests that were retained underwent changes in item content, administration procedures, and scoring procedures. Five new subtests were developed: Picture Concepts, Letter-Number Sequencing, Matrix Reasoning, Word Reasoning, and Cancellation. Of the 15 subtests, 10 are considered core subtests and contribute equally to the FSIQ.

As a part of the standardization process, the WISC-III was compared to the WISC-IV. Both instruments were given to a sample of 244 children ranging in age from 6 to 16. Tests were administered in a counterbalanced order with a mean test-retest interval of 28 days. All mean IQ and factor scores were lower on the WISC-IV than on the WISC-III. Mean scores on the WISC-III for the VCI, POI, FDI, PSI, and FSIQ were 106.0 (*SD* =13.6), 106.9 (*SD* = 14.6), 103.0 (*SD* = 15.9), 108.2 (*SD* = 16.3), and 107.0

(*SD* = 14.4), respectively. Mean scores on the WISC-IV for the VCI, PRI, WMI, PSI, and FSIQ were 102.9 (*SD* = 12.3), 103.9 (*SD* = 14.0), 101.5 (*SD* = 15.3), 102.7 (*SD* = 15.1), and 104.5 (*SD* = 14.0), respectively. Correlation coefficients between the WISC-III and the WISC-IV were .85 (VCI-VCI), .70 (PRI-POI), .74 (WMI-FDI), .81 (PSI-PSI), and .87 (FSIQ-FSIQ) (Wechsler, 2003). To date there have been no studies published examining the comparability between the WISC-III and the WISC-IV other than what was reported in the WISC-IV manual. Refer to Table 4 to see a summary of the WISC-III—WISC-IV comparison study that was included in the WISC-IV manual.

Table 4

Summary of Comparison Study Included in WISC-IV Manual

Study	Sample	Order	Correlations	Mean Differences
Wechsler (2003)	244 children ranged in age from 6 to 16.	Counterbalanced Mean test-retest interval of 28 days	.85 (VCI-VCI) .70 (PRI-POI) .74 (WMI-FDI) .81 (PSI-PSI) .87 (FSIQ-FSIQ)	All mean IQ and factor scores were lower on the WISC-IV than on the WISC-III Mean scores on the WISC-IV for the VCI, PRI, WMI, PSI, and FSIQ were 102.9 (*SD* = 12.3), 103.9 (*SD* = 14.0), 101.5 (*SD* = 15.3), 102.7 (*SD* = 15.1), and 104.5 (*SD* = 14.0), respectively Mean scores on the WISC-III for the VCI, POI, FDI, PSI, and FSIQ were 106.0 (*SD* =13.6), 106.9 (*SD* = 14.6), 103.0 (*SD* = 15.9), 108.2 (*SD* = 16.3), and 107.0 (*SD* = 14.4), respectively

In examining all of the studies comparing versions of the WISC, correlations were significant, typically in the .80s with some that reached the .90 level of significance. Mean differences between the two versions were generally significant as well, with

scores on the newer version being lower. These differences were indicative of the Flynn effect and gave evidence as to why it is essential to renorm intelligence tests every 10 to 15 years.

The present study examined the correlations and mean differences between revised versions of the WISC. Specifically, the purpose of this study was to compare performance differences between IQ scores, factor index scores and verbal-nonverbal difference scores and examine evidence of the Flynn effect between the WISC-III and the WISC-IV in a sample of special education students. IQs and factor scores were compared because they are the main scores computed from an intelligence test and it is imperative to see how they are related between versions of a test. Verbal-nonverbal discrepancies were also examined because at times examiners look for a discrepancy between the two in order to identify similarly or dissimilarly developed abilities. For example, better developed nonverbal skills could be indicative of speech related problems, but better developed verbal skills could demonstrate a nonverbal learning disability.

Research Questions and Hypotheses

Based on previous studies of WB-I—WISC, WISC—WISC-R, WISC-R—WISC-III, and WISC-III—WISC-IV comparisons, this study had two research questions and hypotheses. The first research question was how are WISC-III and WISC-IV IQs, factor index scores, and verbal-nonverbal difference scores related. It was hypothesized that correlations between the WISC-III and the WISC-IV would be in the moderate to high ranges. The second research question examined whether mean differences between similar scores on the WISC-III and WISC-IV were statistically significant and indicative of the Flynn effect. It was hypothesized that scores would be lower on the WISC-IV than

on the WISC-III. See Table 5 for a list of scores that were examined for correlations and significant mean differences.

Table 5

Comparisons Between the WISC-III and the WISC-IV in the Current Study

WISC-III	WISC-IV
Full Scale IQ	Full Scale IQ
Verbal IQ	Verbal Comprehension Index
Performance IQ	Perceptual Reasoning Index
Verbal Comprehension Index	Verbal Comprehension Index
Perceptual Organization Index	Perceptual Reasoning Index
Freedom from Distractibility Index	Working Memory Index
Processing Speed Index	Processing Speed Index
Verbal IQ-Performance IQ Difference	Verbal Comprehension Index—Perceptual Reasoning Index Difference
Verbal Comprehension Index—Perceptual Organization Index Difference	Verbal Comprehension Index—Perceptual Reasoning Index Difference

Method

Participants

Participants in the current study included 89 students with Individualized Education Plans from kindergarten through grade 11 in a medium sized school district in the Midwest. The majority of students were male ($n = 56$). Most of the students in the sample were Black/African American (60.7%), followed by 28.1 % Caucasian/White, 2.2% Hispanic, and 7.9% Other. All students were referred for reevaluation due to state and federal requirements of triennial reevaluations for special education students. The time between evaluations ranged from 12 months to 95 months with an average of 38.87 months. Approximately one third of the students lived with their mother only (34.8%) and another third lived with both parents (25.8%). The rest of the students had family situations such as father only, blended families, and extended families. Students had a variety of disabilities: Specific Learning Disability (SLD), Speech or Language Impairment (SpL), Mental Retardation (MR), Serious Emotional Disturbance (SED) Other Health Impairment (OHI), Autism, Hearing Impairment (HI), Behavior/Emotional Disturbance (BED), Physical Impairment (PI), and Traumatic Brain Injury (TBI). The sample included students who had been in special education at some time between the years 1991 and 2008. As of 2008, total student enrollment of the district was 9,326. Almost half of the student population was Caucasian/White (45.6%), 38.0% of students were African American/Black, 6.7% of students were Latino/Hispanic, 9.4% of students were Asian/Pacific Islander , and 0.3% of students were classified as Other. The district's low income rate was 43.9% and the limited English proficiency rate was 4.3%.

Procedure

Data in the current study were collected through record reviews of students who were receiving special education services or students who had been referred for a psychological evaluation due to a suspected disability. Demographic information (age, gender, race/ethnicity, family status, and primary and secondary languages) was obtained from each student's special education file. Permission to access this information was granted by the School Board of the district and the Special Education Coordinator.

School Psychology graduate students and a professor entered data from special education files onto data coding forms. Trained undergraduates entered data from coding sheets into a computerized spreadsheet. The information included: identification number, evaluation number, date of birth, date of psychological evaluation, chronological age, school attended, grade, sex, race/ethnicity (parent designated), family status, child's language, language spoken in the home environment, retention in grade, school psychologist, disability, cognitive test used during the evaluation, cognitive test composites and factor scores, cognitive subtest scores, achievement tests used during the evaluation, achievement composite scores, achievement subtest scores, adaptive behavior tests used during the evaluation, visual-perceptual-motor tests used during the evaluation, and objective and projective psychopathology tests used during the evaluation. The students' names were not used; the identification number was used to reference students.

Instruments

The WISC-III was an instrument designed to measure cognitive abilities of children aged 6-0 to 16-11. It included 13 subtests, 12 which formed the basis for 4 factor-based scores, and 3 IQ scores. The four factors included: the Verbal

Comprehension Index (VCI), measured by the Information, Similarities, Vocabulary, and Comprehension subtests; the Perceptual Organization Index (POI), measured by the Picture Completion, Picture Arrangement, Block Design, and Object Assembly subtests; the Freedom from Distractibility Index (FDI), measured by the Arithmetic and Digit Span subtests; and the Processing Speed Index (PSI), measured by the Coding and Symbol Search subtests. Although Symbol Search and Digit Span were necessary to compute their respective factor scores, they were not used in calculating the VIQ, PIQ, or FSIQ (Wechsler, 1991).

The WISC-IV is the current version of the Wechsler Intelligence Scale for Children, and like its predecessor, it measures cognitive abilities among children 6-0 through 16-11. It includes 15 subtests, 4 factor scores, and an overall Full Scale IQ (FSIQ). The core battery consists of 10 subtests, all of which are included in deriving the FSIQ and the four factors. The four factors are the Verbal Comprehension Index (VCI), the Perceptual Reasoning Index (PRI), the Working Memory Index (WMI), and the Processing Speed Index (PSI). The VCI is measured by the Similarities, Vocabulary, and Comprehension subtests. The PRI is measured by the Block Design, Picture Concepts, and Matrix Reasoning subtests. The WMI is measured by the Digit Span and Coding subtests. The PSI is measured by the Letter-Number Sequencing and Symbol Search subtests. The five supplemental subtests can be used to replace a subtest from the core battery in computing one of the factor scores or the FSIQ when a subtest is spoiled. Information and Word Reasoning are associated with the VCI, and Picture Completion is related to the PRI. Arithmetic can be replaced for a subtest in the WMI, and the Cancellation subtest can be used in calculating the PSI (Wechsler, 2003).

Both instruments were standardized on samples of 2,200 children who were chosen to closely match the U.S. census data (at the time of the standardization) on age, gender, geographic region, race/ethnicity, and socioeconomic status. On both the WISC-III and the WISC-IV, the four factor scores and the IQs are expressed as standard scores ($M = 100$, $SD = 15$). The VCI subtests generally have the highest *g* loadings, followed by the PRI (POI on the WISC-III), WMI, and PSI (FDI on the WISC-III) subtests. The IQs and factor index reliabilities are usually high (in the .90s), and the subtest reliabilities are generally medium (in the .80s) (Wechsler, 1991; Wechsler, 2003).

Criteria

In order to be included in the current study, participants must have undergone testing for a triennial reevaluation for special education. Additionally, the WISC-III must have been administered at Time 1 and the WISC-IV at Time 2.

Data Analyses

IQs, factor index scores, and verbal-performance difference scores from the WISC-III and the WISC-IV were examined using Pearson product moment correlation analysis. Dependent *t*-tests for differences between means were also conducted to examine the significance between IQs, factor index scores, and verbal-performance difference scores on the WISC-III and the WISC-IV. Cohen's *d* effect size estimate for mean differences was used and interpreted with his criteria: a small effect size was 0.2, a medium effect size was 0.5, and a large effect size was 0.8 (Cohen, 1988). A Cohen's *d* repeated measures calculator that takes into account the correlation between scores was used to determine effect size (Tutorial). The results from this procedure provided information about differences between WISC-III and WISC-IV scores.

Results

Pearson product moment correlations were calculated to determine relationships between similar scores on the WISC-III and the WISC-IV and are presented in Table 6. A correlation of .83 was obtained between WISC-III and WISC-IV FSIQ ($p < .001$). Since the WISC-IV does not include VIQ and PIQ, the WISC-III VIQ and PIQ were compared to the WISC-IV VCI and PRI. A correlation of .75 ($p < .001$) was obtained between the WISC-III VIQ and WISC IV VCI, and a correlation of .65 ($p < .001$) was obtained between the WISC-III PIQ and WISC-IV PRI. Correlations between the WISC-III VCI—WISC-IV VCI ($r = .72$), WISC-III POI—WISC-IV PRI ($r = .70$), WISC-III FDI—WISC-IV WMI ($r = .59$), and WISC-III PSI—WISC-IV PSI ($r = .62$) were statistically significant ($p < .001$). Two verbal-nonverbal difference comparisons were examined. The correlation between the WISC-III VIQ-PIQ difference and the WISC-IV VCI-PRI difference was .17 ($p > .001$). The correlation between the WISC-III VCI-POI difference and the WISC-IV VCI-PRI difference was .20 ($p > .001$). WISC-III and WISC-IV verbal and nonverbal discrepancies were not significantly correlated.

Descriptive statistics for the WISC-III and WISC-IV composite scores are presented in Table 6. A dependent *t*-test revealed a statistically significant mean difference of 3.89 points between WISC-III FSIQ ($M = 83.55$, $SD = 17.26$) and WISC-IV FSIQ ($M = 79.66$, $SD = 14.96$), $t(82) = 3.65$, $p < .001$, $d = .58$. Since the WISC-IV no longer includes VIQ and PIQ, the WISC-III VIQ and PIQ were compared to WISC-IV VCI and PRI, respectively. Dependent *t*-tests revealed no statistically significant differences between the WISC-III VIQ and the WISC-IV VCI, $t(88) = .08$, $p > .05$, $d = .01$ or between the WISC-III PIQ and the WISC-IV PRI, $t(86) = -.48$, $p > .05$, $d = -.07$.

Mean differences between WISC-III indexes and WISC-IV indexes were also examined. Dependent *t*-tests revealed no statistically significant differences between the WISC-III and WISC-IV VCI, $t(81) = 1.55, p > .05, d = .24$; the WISC-III POI and the WISC-IV PRI, $t(64) = .24, p > .05, d = .04$; or the WISC-III FDI and the WISC-IV WMI, $t(77) = 1.09, p > .05, d = .18$. A dependent *t*-test revealed a statistically significant difference between the WISCIII PSI ($M = 91.77$, $SD = 15.04$) and the WISC-IV PSI ($M = 85.15$, $SD = 11.33$), $t(43) = 3.66, p < .001, d = .81$. Mean differences between verbal-performance differences on the WISC-III and the WISC-IV were also examined. Dependent *t*-tests revealed no statistically significant differences between the WISC-III VIQ-PIQ and the WISC-IV VCI-PRI verbal-nonverbal differences, $t(86) = 0.66, p > .05, d = .10$ or between the WISC-III VCI-POI and the WISC-IV VCI-PRI verbal-nonverbal differences, $t(64) = 1.81, p > .05, d = .32$. All mean differences represented small effect sizes except for a medium effect size for FSIQ and a large effect size for PSI.

Table 6

Descriptive Statistics, Pearson Product Moment Correlations, Dependent t-tests, and Effect Size Estimates for WISC-III and WISC-IV Scores

			WISC-III		WISC-IV			
	n	*r*	*M*	*SD*	*M*	*SD*	*t*	*d*
FSIQ	83	.83*	83.55	17.26	79.66	14.96	3.65*	0.58
VIQ/VCI	89	.75*	84.38	17.21	84.28	16.09	.08	0.01
PIQ/PRI	87	.65*	84.97	16.85	85.68	16.08	-.48	-0.07
VCI	82	.72*	85.67	16.04	83.67	15.18	1.55	0.24
POI/PRI	65	.70*	85.97	17.01	85.58	15.54	.24	0.04
FDI/WMI	78	.59*	82.87	13.05	81.33	14.19	1.09	0.18
PSI	44	.62*	91.77	15.04	85.16	11.33	3.66*	.81
VP Diff 1	87	.17	.05	12.05	-1.09	12.98	.66	.10
VP Diff 2	65	.20	1.60	13.00	-2.08	13.02	1.81	.32

Note. FSIQ = Full Scale IQ, VIQ = Verbal IQ, PIQ = Performance IQ, VCI = Verbal Comprehension Index, PRI = Perceptual Reasoning Index, POI = Perceptual Organization Index, FDI = Freedom From Distractibility Index, WMI = Working Memory Index, PSI = Processing Speed Index, VP Diff 1 = Verbal-Performance Difference (WISC-III VIQ-PIQ, WISC-IV VCI-PRI), VP Diff 2 = Verbal-Performance Difference (WISC-III VCI-POI, WISC-IV VCI-PRI), *r* = Pearson product moment correlations *d* = Cohen's *d* effect size estimate (Cohen, 1988)

*$p < .001$

Discussion

The present study examined relationships between the WISC-III and the WISC-IV and compared mean performance differences for evidence of the Flynn effect in a sample of special education students. Generally, previous studies of WB-I—WISC (Delattre & Cole, 1952; Price & Thorne, 1955), WISC—WISC-R (Hamm et al., 1976; McGonagle, 1977; Reschley & Davis, 1977; Spitz, 1983; & Swerdlik 1978), WISC-R—WISC-III (Bolen, Aichinger, Hall, & Webster, 1995; Hamm et al., 1996; Lyon, 1995; Slate & Saarnio, 1995; and Wechsler, 1991), and WISC-III—WISC-IV (Wechsler, 2003) comparisons found moderate to high correlations between the older and newer WISC for FSIQ, VIQ, and PIQ as well as statistically significant mean differences between the older and newer versions of the WISC, with scores being lower on the newer version.

The first research question in the present study examined the correlations between IQs, factor index scores, and verbal-performance difference scores on the WISC-III and the WISC-IV. It was hypothesized correlations would be in the moderate to high ranges. Findings from the current study demonstrated the WISC-III FSIQ was highly correlated with the WISC-IV FSIQ. These results were consistent with findings of Delattre and Cole (1952), Price and Thorne (1955), Hamm et al. (1976), Swerdlik (1978), Reschly and Davis (1977), Spitz (1983), McGonagle (1977), Wechsler (1991), Bolen et al. (1995), Slate and Saarnio (1995), Lyon (1995), and Wechsler (2003). This finding was much anticipated. Of all correlations examined in this study and previous studies, the FSIQ should have had the highest correlation because the FSIQ is considered the most stable score on an IQ test and does not generally change over time.

Due to the nature of the WISC-IV revision, Wechsler (2003) is the only previous study that also examined verbal and performance scores by correlating the WISC-III VIQ with the WISC-IV VCI (r = .83) and the WISC III PIQ with the WISC-IV PRI (r = .73). Results of the current study were consistent with Wechsler, finding high but somewhat lower correlations. It is possible the current study obtained lower correlations than Wechsler due different samples and more control in the Wechsler study. The elimination of the VIQ and PIQ from the WISC-IV was based on new research and theory that no longer supported the dual IQ model. Although those scores are no longer available, findings from the current study and Wechsler suggested that a good estimate of verbal and/or nonverbal performance can still be obtained on the WISC-IV without the VIQ and PIQ. These results were supportive of the new framework for the WISC-IV.

The current study investigated the relationship between the four factor scores on the WISC-III and the WISC-IV, and although the WISC-III initially introduced the four factor index scores, only one previous study (Wechsler, 2003) also examined the correlations between the four factor scores. Results from both studies indicated factor index scores on the WISC-III and the WISC-IV were highly correlated. These correlations gave support to the similarity in content between the WISC-IV and the WISC-III. Although there were several content changes, the two tests still measure similar constructs. Two of the factors were renamed on the WISC-IV (POI to PRI and FDI to WMI) but they still measure similar abilities. Perhaps, the new names on the WISC-IV are more indicative of the types of skills they measure. Even though the current study and Wechsler's study indicated highly correlated factor scores, three of the correlations were higher in the Wechsler study (VCI, FDI/WMI, PSI). It is possible that

the correlations were stronger in the Wechsler study as a result of more control in the study and a more representative sample of the general population, instead of a sample of students with disabilities like the current study.

Weiner and Kaufman (1979) was the only previous study that examined the verbal-nonverbal difference scores, but correlations were not reported so there are no previous studies to which results from the current study could be compared. In spite of the hypothesis that all scores would be significantly correlated, the lack of correlation between verbal-performance difference scores was not surprising. While composite scores are fairly stable, the difference between two composite scores would have less stability and should not be expected to be consistent. Similar stability results were obtained in Canivez and Watkins (1998; 1999; 2001). Their results demonstrated the WISC-III VCI and WISC-III POI were stable over a three-year period, but the WISC-III VCI-POI difference was not.

The second research question examined the mean differences between similar scores on the WISC-III and the WISC-IV for evidence of the Flynn effect. It was hypothesized that the mean differences between the IQs, factor index scores, and verbal-nonverbal difference scores would be statistically significant. When examining the mean difference between the WISC-III FSIQ and the WISC-IV FSIQ, results of this study were comparable to previous WISC—WISC-R studies (Spitz, 1983) and WISC-R—WISC-III studies (Hamm et al., 1996). There was a difference of 5 points in Spitz and 5 points in Hamm et al. Like the current study, these studies used a reevaluation format with similar samples, so it makes sense that the results would be similar. The results of the current study were also similar to the Wechsler (2003) WISC-III—WISC-IV study despite the

study using a counterbalanced order with a mean test-retest interval of only 28 days. A difference of approximately three points was found by Wechsler. Although Wechsler and the current study both compared the WISC-III to the WISC-IV, there was the added factor of a time delay in the current study. The WISC-IV FSIQ was significantly lower than the WISC-III FSIQ as would be expected based upon evidence of the Flynn effect (Flynn, 1984, 1987). After a test has been around for more than 10 years, scores tend to become inflated, giving a deceptive view of one's true intellectual functioning.

As with the examination of correlations, only one previous study examined the mean difference between the WISC-III VIQ and the WISC-IV VCI and the WISC-III PIQ and the WISC-IV PRI (Wechsler, 2003). Similar to Wechsler, results of the current study indicated no significant differences between the WISC-III VIQ and WISC-IV VCI or between the WISC-III PIQ and the WISC-IV PRI. Although the VIQ and VCI and PIQ and PRI are similar in content, the findings of both the current study and Wechsler could be due the fact the scores are different enough from one another that mean differences would not be identified. Factor analysis has shown that the VIQ and the PIQ contain subtests that are not verbally or perceptually oriented.

Similar to the comparison study included in the WISC-IV manual, the current study also examined mean differences between factor index scores. Unlike results from the WISC-III—WISC-IV study in the WISC-IV manual, this study found only one significant mean difference between the WISC-IV and WISC-III factor index scores. The WISC-III—WISC-IV PSI comparison provided evidence for the Flynn effect. According to Wechsler (2003), in comparison to the other factors, the PSI was not substantially changed during the revision, which would account for finding the expected mean

difference. Since the WISC-III PSI is measuring similar skills as the WISC-IV PSI, it makes sense that the score on the WISC-III would be an inflated representation of a child's processing abilities. Although a smaller sample size makes it harder to show significant differences, perhaps there was less variation in the scores; only 44 participants had PSI scores on both the WISC-III and the WISC-IV. It is possible that this study did not find significant mean differences between all four factor indexes like Wechsler because the two studies used different formats (the current study used a reevaluation format with a time delay and Wechsler used a counterbalanced approach with only a 28 day test-retest interval) with dissimilar samples (Wechsler adequately represented the general population and the current study was heavily loaded with students with mental retardation and specific learning disability). Limited significant mean differences between factor scores could also be due to the structural changes made on the factor scores. As mentioned in the WISC-IV manual, the WMI and PRI underwent drastic changes during the revision, more than previous revisions of the WISC (Wechsler, 2003). Since there are no other studies that examine the mean differences between factor scores, it is hard to say what other studies would find and whether the results of the current study are atypical, but replication will help determine this.

Weiner and Kaufman (1979) is the only previous study that examined the relationship between verbal-nonverbal difference scores between versions of the WISC. Weiner and Kaufman compared the WISC VIQ-PIQ difference to the WISC-R VIQ-PIQ difference and found a mean difference of 0.2, which was not significant. Results of the current study were similar to Weiner and Kaufman. Neither comparison in the current study (WISC-III VIQ-PIQ—WISC-IV VCI-PRI or WISC-III VCI-POI—WISC-IV VCI-

PRI) yielded significant mean differences. Findings from the current study as well as Weiner and Kaufman suggest cautious examination of the verbal-performance difference scores because what a student obtains at Time 1 may not be reflected in his or her score at Time 2. Despite composite scores being fairly stable, the difference between two composite scores would have less stability and should not be expected to be consistent. Canivez and Watkins (1998; 1999; 2001) demonstrated similar stability results, which indicated the WISC-III VCI and WISC-III POI to be stable over a three-year period, but not the WISC-III VCI-POI difference.

The decrease in scores on the FSIQ and PSI may be reflective of renorming and content changes. Although not all of the score comparisons were significantly different, this is not evidence to disconfirm the Flynn effect; the results just do not add to the already existing evidence of the Flynn effect. The lack of significant results is due to the limitations of the study.

Limitations

As with all research studies, this study had several limitations. The sample is a major limitation of this study. First, the sample was small. Additionally, the sample was not representative of the overall population of the school district from which the data were taken or the general population; the sample was comprised mainly of African American/Black males receiving special education services under Specific Learning Disability or Mental Retardation. This made it hard to generalize the results to other populations. Another limitation of the study was the record review format. Because it was a record review, there were no true controls for internal validity in terms of administration. Since this study was conducted as a record review, there were no set

criteria for which scores needed to be reported. While going through the files, it was evident that some scores were not computed. Error in data input is also a possibility due to several different data collectors. The data collection for this study took place over a period of two years, so it is possible that more files could have been reviewed if more time had been allotted for this aspect of the study. Another limitation of this study was the large range in time between test administrations, which ranged from 12 months to 95 months. This is problematic because it does not allow for internal consistency in the study, nor external validity as reevaluations occur every three years. Although the WISC-III and the WISC-IV are different tests, they are comprised of similar content and activities. Research shows practice effects tend diminish within a year, but it is possible those students who were tested right at a year may have experienced minimal practice effects (Kaufman, 1994). Another limitation of the study is that for 75% of the students, the WISC-IV was administered by a different psychologist than the WISC-III. Although this is typical in the school setting, parts of the WISC are subjective in scoring, and this could result in different scores. According to Erdodi, Richard, and Hopwood (2009), the subtests that require subjective scoring on the WISC-IV are the most prone to errors, especially when students have low ability levels. In this study it is possible that psychologists over- or underestimated students' abilities due to several students in the sample having mental retardation.

Future Research

The results of this study lead to suggestions for future research. Simply replicating this study would be beneficial. To date, there are still few published studies comparing the WISC-IV to the WISC-III. Since replication would likely require a record

review, changes such as including only students who have all scores or using only one or two data collectors would be an improvement. Future studies should attempt to use a more heterogeneous sample of students in order to make the results more generalizable. Dividing the sample into different groups (race or disability) might provide a good comparison for future studies as well. Further examining relationships between the WISC-III VIQ and the WISC-IV VCI as well as the WISC-III PIQ and the WISC-IV PRI would be beneficial. Looking at the relationship between WISC-III and WISC-IV verbal-performance difference scores might be advantageous because it would provide new information since the study in WISC-IV manual did not even compare these scores.

Implications

Based on the results of this study, there are several suggestions for practicing school psychologists. Although a Response to Intervention (RTI) format is becoming one mandated approach for determining Specific Learning Disability eligibility in some areas, standardized assessment is still a main approach used in some states and it will always have its place in school psychology regardless of eligibility criteria. When comparing scores between the WISC-III and the WISC-IV practitioners can be assured that the two tests cover similar content, considering high correlations between most of the scores on the two versions of the WISC. As far as the relationship between verbal-performance difference scores, that should be interpreted with caution due to the difference in structure between the WISC-III and the WISC-IV. The WISC-III verbal-nonverbal difference score may be representative of different skills than the WISC-IV verbal-nonverbal difference score. Additionally, studies examining test-retest stability of the WISC-III verbal-performance difference show poor agreement across time. Canivez and Watkins

(1998) found variation in VIQ-PIQ discrepancies to range from ± 10 points to as much as 45 points. The results of the current study demonstrated the same phenomenon when comparing the WISC-III verbal-nonverbal difference to the WISC-IV verbal-nonverbal difference. In looking at the mean differences between the WISC-III and the WISC-IV, the results of this study lead one to the conclusion that the Flynn effect exists, with respect to the FSIQ and the PSI. When comparing those two scores on the WISC-III and the WISC-IV, it should be noted that the scores on the WISC-III are inflated at the last years of its use and not a good estimate of a child's true intellectual functioning. Such findings suggest that it is beneficial to renorm tests of cognitive abilities more frequently.

References

Adams, K. M. (2000). Practical and ethical issues pertaining to test revision. *Psychological Assessment, 12*(3), 281-286.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington DC: American Psychological Association.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington DC: Author.

Bolen, L. M., Aichinger, K. S., Hall, C. W., & Webster, R. E. (1995). A comparison of the performance of cognitively disabled children on the WISC-R and the WISC-III. *Journal of Clinical Psychology, 51*(1), 89-94.

Canivez, G. L., & Watkins, M. W. (2001). Long-term stability of the Wechsler Intelligence Sale for Children-Third Edition among students with disabilities. *School Psychology Review, 30*(3), 438-453.

Canivez, G. L., & Watkins, M. W. (1999). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition among demographic subgroups: Gender, race/ethnicity, and age. *Journal of Psychoeducational Assessment, 17*, 300-313.

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children. *Psychological Assessment, 10*(3), 285-291.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Delattre, L., & Cole, D. (1952). A comparison of the WISC and the Wechsler-Bellevue. *Journal of Consulting Psychology, 16*(3), 228-230.

Erdodi, L. A., Richard, D. C. S., & Hopwood, C. (2009). The importance of relying on the manual. *Journal of Psychoeducational Assessment, 27*(5), 374-385,

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Gironda, R. J. (1977). A comparison of WISC and WISC-R results of urban educable mentally retarded students. *Psychology in the Schools, 14*(3), 271-275.

Goh, D. S., Teslow, C. J., & Fuller, G. B. (1991). The practice of psychological assessment among school psychologists. *Professional Psychology, 12*(6), 696-706.

Hamm, H., Wheeler, J., McCallum, S., Herrin, M., Hunter, D., & Catoe, C. (1976). A comparison between the WISC and the WISC-R among educable mentally retarded students. *Psychology in the Schools, 13*(1), 4-8.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd ed.). Columbus, OH: Merrill.

Hutton, J. B., Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review, 21*(2), 271-284.

Kaufman, A. S. (1994). Practice effects. In R. Sternberg (Ed.), *Encyclopedia of human intelligence* (Vol. 2, pp. 828-833). New York: Mcmillan.

Lyon, M. A. (1995). A comparison between WISC-III and WISC-R scores for learning disabilities reevaluations. *Journal of Learning Disabilities, 28*(4), 253-255.

Mahone, E. M., Miller, T. L, Koth, C. W., Mostofsky, S. H., Goldberg, M. C., & Denckla, M. B. (2003). Differences between WISC-R and WISC-III performance scale among children with ADHD. *Psychology in the Schools, 40*(4), 331-340.

McGinley, P. (1981). A comparison of WISC and WISC-R test results. *The Irish Journal of Psychology, 1*, 23-24.

McGonagle, B. (1977). A comparison between the WISC and the WISC-R among a clinical referred population. *Psychology in the Schools, 14*(4), 423-426.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–181). New York: Guilford Press.

Moffitt, T. E., Caspi, A., Harkness, A. R., Silva, P.A. (1993). The natural history of change is intellectual performance: Who changes? How much? Is it meaningful? *Journal of Child Psychology and Psychiatry, 34*, 455-506.

Price, J. R., & Thorne, G. D. (1955). A statistical comparison of the WISC and the Wechsler-Bellevue, Form I. *Journal of Counseling Psychology, 19*(6), 479-482.

Reschly, D. J., & Davis, R. A. (1977). Comparability of WISC and WISC-R scores among borderline and mildly retarded children. *Journal of Clinical Psychology, 33*(4), 1045-1048.

Sattler, J. M. (2001). *Assessment of children cognitive applications* (4th ed.). San Diego: Jerome M. Sattler, Publisher, Inc.

Schwarting, F. G. (1976). A comparison of the WISC and WISC-R. *Psychology in the Schools, 13*(2), 139-141.

Sevier, R. C., & Bain, S. K. (1994). Comparison of WISC-R and WISC-III for gifted students. *Roeper Review, 17*(1), 39-42.

Shalizi, C. (2009). The domestication of the savage mind. *American Scientist, 97*, 244-247.

Slate, J. R., & Saarnio, D. A. (1995). Differences between WISC-III and WISC-R IQs: A preliminary investigation. *Journal of Psychoeducational Assessment, 13*, 340-346.

Solly, D. C. (1977). Comparison of WISC and WISC-R scores of mentally retarded and gifted children. *Journal of School Psychology, 15*(3), 255-258.

Solway, K. S., Fruge, E., Hays, J. R., Cody, J., & Gryll, S. (1976). A comparison of the WISC and WISC-R in a juvenile delinquent population. *The Journal of Psychology, 94*, 101-106.

Spitz, H. H. (1983). Intratest and intertest reliability and stability of the WISC, WISC-R, and WAIR full scale IQS in a mentally retarded population. *The Journal of Special Education, 17*(1), 69-80.

Steen, R. G. (2009). Are people getting smarter? *Human intelligence and medical illness* (1-8). New York: Springer New York.

Stinnett, T. A., Havey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12*(4), 331-350.

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12*(3), 237-244.

Swerdlik, M. E. (1978). Comparison of WISC and WISC-R scores of referred black, white and latino children. *Journal of School Psychology, 16*(2). 110-125.

Tutorial 5 advanced ANOVA, power and effect sizes. Retrieved January 25, 2010, from the Wilderdom website: http://wilderdom.com/courses/surveyresearch/tutorials/5/

Vance, H., Maddux, C. D., Fuller, G. B., & Awadh, A. M. (1996). A longitudinal comparison of the WISC-III and WISC-R scores of special education students. *Psychology in the Schools, 33*, 113-118.

Watkins, C. E., Jr., Campbell, V. L., Nierberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*(1), 54-60.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children Revised (WISC-R)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children Third Edition (WISC-III).* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Manual for the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV).* San Antonio, TX: The Psychological Corporation

Weiner, S. G., & Kaufman, A. S. (1979). WISC-R versus WISC for black children suspected of learning or behavioral disorders. *Journal of Learning Disabilities, 12*(2), 41-46.