8-13-2019

Structural Validity Evidence of the Wechsler Intelligence Scale for Children- Fifth Edition with African-American Students who have been Referred for Evaluation

Rachel Y. Taylor
Georgia State University

Follow this and additional works at: https://scholarworks.gsu.edu/eps_diss

ACCEPTANCE

This dissertation, STRUCTURAL VALIDITY EVIDENCE OF THE WECHSLER

INTELLIGENCE SCALE FOR CHILDREN- FIFTH EDITION WITH AFRICAN-AMERICAN

STUDENTS WHO HAVE BEEN REFERRED FOR EVALUATION, by RACHEL YVONNE

TAYLOR, was prepared under the direction of the candidate's Dissertation Advisory

Committee. It is accepted by the committee members in partial fulfillment of the requirements

for the degree, Doctor of Philosophy, in the College of Education and Human Development,

Georgia State University.

The Dissertation Advisory Committee and the student's Department Chairperson, as

representatives of the faculty, certify that this dissertation has met all standards of excellence and

scholarship as determined by the faculty.

_____   _____
C. Kevin Fortner, Ph.D.   Tiffany Hogan, Ph.D.
Committee Chair   Committee Member

_____   _____
William Curlette, Ph.D.   Hongli Li, Ph.D.
Committee Member   Committee Member

_____
Date

_____
William Curlette, Ph.D.
Chairperson, Department of Educational Policy
Studies

_____
Paul Alberto, Ph.D.
Dean, College of Education & Human Development

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced

degree from Georgia State University, I agree that the library of Georgia State University shall

make it available for inspection and circulation in accordance with its regulations governing

materials of this type. I agree that permission to quote, to copy from, or to publish this

dissertation may be granted by the professor under whose direction it was written, by the College

of Education and Human Development's Director of Graduate Studies, or by me. Such quoting,

copying, or publishing must be solely for scholarly purposes and will not involve potential

financial gain. It is understood that any copying from or publication of this dissertation which

involves potential financial gain will not be allowed without my written permission.

_____
Rachel Yvonne Taylor

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance

with the stipulations prescribed by the author in the preceding statement. The author of this

dissertation is:

Rachel Yvonne Taylor
Educational Policy Studies
College of Education and Human Development
Georgia State University

The director of this dissertation is:

C. Kevin Fortner, Ph.D.
Department of Educational Policy Studies
College of Education and Human Development
Georgia State University
Atlanta, GA 30303

CURRICULUM VITAE

Rachel Taylor

ADDRESS: 2041 Millstone Drive, S.W.
                           Conyers, GA 30094

EDUCATION:

Doctor of Philosophy	2019	Georgia State University Research, Measurement, and Statistics
Education Specialist	2000	Valdosta State University School Psychology
Master of Education	1997	Valdosta State University School Counseling
Bachelor of Business Administration	1989	Georgia State University Management

PROFESSIONAL EXPERIENCE:

2004-Present   School Psychologist
                           Atlanta Public Schools
                           Atlanta, GA

2003-2004   School Psychologist
                           Fairfax County Public Schools
                           Alexandria, VA

2002-2003   School Psychologist
                           Charleston County Public Schools
                           Charleston, SC

PROFESSIONAL SOCIETIES AND ORGANIZATIONS
        2004   National Association of School Psychologists
        1997   Georgia Association of School Psychologists

STRUCTURAL VALIDITY EVIDENCE OF THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN- FIFTH EDITION WITH AFRICAN-AMERICAN STUDENTS WHO HAVE BEEN REFERRED FOR EVALUATION

by

RACHEL TAYLOR

Under the Direction of Dr. C. Kevin Fortner

ABSTRACT

The Wechsler Intelligence Scale for Children- Fifth Edition (WISC-V; Wechsler, 2014) is expected to be the most widely used measure of intelligence of school aged children in the United States for at least the next ten years. Results of this assessment are used to make decisions about students' educational placements. Evidence of structural validity of previous versions of the Wechsler scales with subjects who have been referred for evaluation due to a suspected disability was rarely examined. Only six studies focused on evidence of structural validity of the Wechsler Intelligence Scale for Children- Fourth Edition with referred samples, during its reign as the leading measure of intelligence for over a decade. In five of those studies, researchers employed confirmatory factor analysis, while exploratory factor analysis was employed in one study. In this study, I investigated the factor structure and measurement invariance of the WISC-V with students who had been referred for evaluation because of academic and/or behavioral

difficulties. Participants were African-American students in one urban school district in the southeastern United States who were administered the WISC-V during the 2015-2017 school years. Confirmatory factor analysis was conducted using Mplus version 7.4 to determine whether the referred sample's data fit the 10 primary subtest structural model that is published in the *WISC-V Technical and Interpretive Manual* (Wechsler, 2014). Results indicated that the best fitting structural model was the four-factor hierarchical model, not the five-factor hierarchical model that the publishers endorse. Measurement invariance of the WISC-V between genders was also investigated using Mplus version 7.4. Invariance between genders was confirmed with the four-factor model. The five-factor model of the female sample's data would not converge, which suggested measurement variance between genders or one of several other problems commonly associated with small sample sizes and large numbers of parameters to estimate. Interpretation of the WISC-V should be based on the Full Scale IQ only. Recommendations for future research were offered.

STRUCTURAL VALIDITY EVIDENCE OF THE WECHSLER INTELLIGENCE SCALE
FOR CHILDREN- FIFTH EDITION WITH AFRICAN-AMERICAN STUDENTS WHO
HAVE BEEN REFERRED FOR EVALUATION

by

RACHEL TAYLOR

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

Research, Measurement, and Statistics

in

Educational Policy Studies

in

the College of Education and Human Development

Georgia State University

Atlanta, GA
2019

DEDICATION

This dissertation is dedicated to my father, Willie James Taylor, who departed this life too soon,

but not before doing great things for our family. He instilled in us an unwavering sense of

determination. I will always be grateful that he was my father.

ACKNOWLEDGMENTS

First, I would like to thank the chair of my dissertation committee, Dr. Charles Kevin Fortner. You have no idea how much your kind words and patience have meant to me throughout this process. You were often my "soft place to land," and I deeply appreciate everything you have done for me. I am grateful for your general ideas about the broad topic to the suggestions about changing specific words in the dissertation. Having you as a mentor has been a privilege and an honor.

I am deeply appreciative of the contributions of my other committee members, Drs. William Curlette, Tiffany Hogan, and Hongli Li. Each of you offered your professional talents to improve this project. Dr. Curlette offered the wisdom garnered through years of leading our department. Dr. Hogan connected the dots to help me develop a smoother flowing document, while Dr. Li assisted me with technical support.

Last, but certainly not least, I would like to thank my family. My mother, Molly Taylor, will always be my inspiration. To my sister, Shonna, and her family, Ernest, Myles, and McKinley, I thank you for your support and encouragement. To my brother, Gayron, and his sons (my nephew sons), Austin and Ryan, I cannot express how much you have enriched my life. I have completed my journey, and I will join the family in all activities!

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Validity of the WISC-V with a Referred Sample

I have been employed as a school psychologist for many years. I am very passionate about my profession and love what I do for children; however, I am often concerned about whether I am making the right decisions about students. The primary responsibility of my profession is to evaluate students who are suspected of having disabilities. Based on results of those evaluations, inferences are made about the students, along with decisions about their educational services. In this study, I will investigate the validity of the *Wechsler Intelligence Scale for Children- Fifth Edition* (WISC-V; Wechsler, 2014). This instrument is expected to be the most widely used measure of intelligence of school aged children in the United States for at least the next ten years.

In Chapter 1, I will provide information about the Wechsler scales and the current version of the assessment. I will discuss the importance of validity and point out that there is little evidence to support the validity of previous versions of this instrument with certain subpopulations. While the publisher, NCS Pearson, Inc., has done a better job of evidencing validity with various disability groups, evidence of structural validity of previous versions of the Wechsler Intelligence Scale for Children with subjects who had been referred for evaluation due to a suspected disability was rarely examined by independent researchers. When it was examined, confirmatory factor analysis was utilized in the vast majority of the studies. This study will be significant, because I will examine evidence of structural validity of the Wechsler Intelligence Scale for Children- Fifth Edition with students who have been referred for evaluation because of academic and/or behavioral difficulties. While separate studies have examined the structural models derived from the standardization sample of this version of the Wechsler scales by employing confirmatory factor analysis (Wechsler, 2014), exploratory factor analysis (Canivez et al., 2014,

2015), and exploratory bifactor analysis (Dombrowski, 2015), no studies that examined the structural validity of the WISC-V with a referred sample have been published. This information will be useful in developing methods of interpreting results of the assessment.

In Chapter 2, I will review the literature on test validation and discuss the publisher's recommendations for interpreting results of the WISC-V. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) specified five sources of validity evidence: evidence based on test content, response processes, testing consequences, internal structure, and relations to other variables. Each source of validation will be discussed, along with how it will be addressed in this study. Of the five types of validity evidence described in the *Standards*, only one of them, internal structure, will be examined in this study. In gathering research on test validation, it was discovered that research by independent investigators does not support the recommended method of interpreting WISC-V results.

In Chapter 3, I will provide detailed information about my research methodology. I will use Mplus version 7.4 to conduct confirmatory factor analyses to determine whether the referred sample's data fit the structural model published in the WISC-V manual. I will also use Mplus version 7.4 to examine measurement invariance between genders on the WISC-V.

In Chapter 4, I will present the results of my study.

In Chapter 5, I will discuss those results, make suggestions about interpreting scores on the WISC-V, and make recommendations about future research on this topic.

1 THE PROBLEM

Wechsler intelligence scales are among the most widely used instruments for measuring intelligence in the world. To date, approximately twenty countries have adapted and standardized

Wechsler intelligence scales. These scales are popular and well-respected because of their psychometric properties and practical relevance. Wechsler scales are frequently used in psychoeducational assessments (Chen, Zhang, Raiford, Zhu, & Weiss, 2015).

The Wechsler Full Scale IQ is used to differentially classify mental disability and giftedness and to identify discrepancies in expected and observed school achievement as related to learning disabilities. It is also used to exclude ability problems in the identification of emotional disturbance and communication disorders. The legitimacy of such claims is entirely dependent on the accuracy of test scores in reflecting individual differences (McDermott, Watkins, & Rhoad, 2014).

Implicit in this practice is the assumption that Wechsler intelligence scale scores have the same meaning for students of different subgroups of the population. Therefore, investigating the measurement invariance of the Wechsler intelligence scales is critical (Chen et al., 2015).

Over the past few decades, there has been a concern about the disproportionality of African-American students identified as needing special education services (Zhang, Katsiyannis, Ju, & Roberts, 2014). There have been consistent findings of overrepresentation of African-American students in special education programs in general, as well as special education programs for students with intellectual disabilities and emotional disturbance. African-American students are the group most overrepresented in special education programs in almost all states. The rates of disproportionality of minority groups tend to increase as the percentage of that minority in the state's population increases. Also, rates of disproportionality are found less frequently in the areas of special education that require less subjectivity in determining eligibility, such as hearing impaired, vision impaired, and orthopedically impaired. Disparities are more

prevalent in areas such as intellectual, emotional, and learning disabilities, in which more judgment is used in determining eligibility. Test bias has not been ruled out as a contributing factor to these disparities (Skiba, 2013).

Invariance is a fundamental property for any clinical instrument that may be used to compare individuals from different groups within a population. Meaningful comparisons can only be made if the measures are comparable. Use of an instrument in making comparisons among groups is supported by evidence of measurement invariance of the instrument. This means that people of the same ability level should be expected to earn the same score, regardless of their race, economic status, or membership in another subgroup of the population (Chen & Zhu, 2012).

For the first 60 years of development of Wechsler scales, the Full Scale IQ, Verbal IQ, and Performance IQ were preserved to provide continuity to its users. Because of this provision of continuity, the test has been criticized for its failure to incorporate modern theories of cognitive abilities. Over time, to incorporate modern theories, the Wechsler Intelligence Scale for Children has progressed from a two-factor to a five-factor instrument.

Published in 2014, the Wechsler Intelligence Scale for Children- Fifth Edition (WISC-V) is the latest version of Wechsler's test of child intelligence. It represents a significant change from the previous version, Wechsler Intelligence Scale for Children- Fourth Edition (WISC-IV), because it incorporates a five-factor scoring model, as opposed to the four-factor model previously used. Since the creation of the Wechsler scales, studies have been conducted on the version of the Wechsler that was currently being used. Over the past decade, studies worldwide have supported the WISC-IV measurement invariance between genders, and across cultures, ages, and

clinical status. Studies of the WISC-IV also found support of the five-factor structure among normative and clinical samples. However, there is little evidence to support the consistency of measurement across subpopulations (Chen et al., 2015).

The *WISC-V Technical and Interpretive Manual* (Wechsler, 2014) states that although test developers are responsible for providing initial evidence of validity, examiners must determine whether evidence supports the use of the test for its intended purpose. A comprehensive evaluation of an instrument's validity evidence includes examination of the relevant literature on previous versions of the instrument, as well as the literature on a newly revised measure for different purposes, in different settings, or with different populations.

The WISC-V *Technical and Interpretive Manual* (Wechsler, 2014) provides evidence of validity for several subgroup populations. Those subgroups include children identified as intellectually gifted, with mild or moderate intellectual disability, with borderline intellectual functioning, with specific learning disorder- reading, with specific learning disorder- reading and written expression, with specific learning disorder-math, with attention deficit/hyperactivity disorder, with disruptive behavior, with traumatic brain injury, who are English language learners, with autism spectrum disorder with language impairment, and with autism spectrum disorder without language impairment. Almost all children in special education will be in one of these groups, so investigating the validity of the WISC-V when applied to these special groups is crucial.

Validity is about the meaning of scores. It measures the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test (AERA et al., 2014). What are validated are the inferences, interpretations, actions, or decisions that are

made based on test scores. Therefore, validity is about the degree to which our inferences are appropriate. For example, in the state of Georgia, a student is considered moderately intellectually disabled, if he has an intelligence quotient of less than 55, along with adaptive scores more than two standard deviations below the mean. Moderately intellectually disabled students are eligible for special education services. Intellectually disabled students remain in the general education setting to the greatest extent possible, but services for students with moderate intellectual disabilities often consist of placement in self-contained classrooms. Instead of meeting the grade level standards outlined by the state, these students must meet the goals of their individualized education plans (IEPs). These goals generally require the students to acquire skills that are below those required for their current grade placement. Therefore, upon graduation, these students do not meet the state requirements to receive a standard diploma. Moderately intellectually disabled students and students with other serious conditions or disabilities receive diplomas that will not allow them to attend college or enlist in the armed forces. This can drastically change the quality of life for these graduates. It would be most unfortunate, if the decisions to qualify these students for special education services were based on inaccurate inferences.

Students belonging to groups that tend to perform less well on the intelligence tests used may be victims of measurement variance. For example, African-American and Hispanic students tend to perform less well on assessments that are verbally loaded or require the respondents to construct verbal responses to inquiries. Their scores on these assessments may be lower than those of other students who show similar levels of intelligence in all other areas (academic, adaptive, social). These students' cultural characteristics may be negatively affecting their performance on those verbally loaded assessments. Therefore, scores of students in these groups would not have the same meaning as those of students in other ethnic groups. The assessments would

be determined to have measurement variance and would not be appropriate for use with these subpopulations.

Throughout the country, school psychologists utilize the Wechsler scales to help determine eligibility for special education services. In analyzing the cause of and working to eradicate the disproportionality of African-American students in special education, practitioners must assess the validity of the Wechsler scales among people who are referred for evaluation, particularly those over represented in the special education population. It is the practitioner's responsibility to determine whether evidence supports the use of the WISC-V and other standardized instruments for their intended purposes (Wechsler, 2014).

The publisher of the Wechsler scales, NCS Pearson, Inc., recommends that scores be used to (a) assess general intellectual functioning; (b) assess performance in each major domain of intelligence; (c) find strengths and weaknesses in each domain of intelligence; (d) interpret clinically relevant score patterns typically found in diagnostic groups; (e) diagnostically and prescriptively interpret the scatter of subtests; (f) make recommendations for teachers to use in class; (g) analyze score profiles from interindividual and intraindividual perspectives; and (h) statistically contrast and interpret differences between pairs of component scores and between individual scores and subsets of multiple scores. School psychologists follow these instructions, make inferences about students, and use those inferences in making educational decisions about students. However, research by independent investigators does not support the use of the WISC-V in making such inferences (McDermott, Watkins, & Rhoad, 2014). Dombrowski, Canivez, Watkins, & Beaujean (2015) and Canivez, Watkins, & Dombrowski (2015, 2017) supported only interpreting the Full Scale IQ (FSIQ).

Research Questions

Does the WISC-V measure the same constructs for a sample of African-American students who have been referred for evaluation as compared to the standardization sample?

Will a confirmatory factor analysis of data gathered from 10 subtests of a referred sample fit the factor structure published in the WISC-V manual?

Is the factor structure of the WISC-V invariant across genders in a referred sample of African-American students?

Does the factor structure derived from a referred sample support interpreting the FSIQ, index scores, and subtest scores of the WISC-V or should interpretations be based on the FSIQ only?

Purpose

The purpose of this study is to investigate evidence of validity based on internal structure of the Wechsler Intelligence Scale for Children- Fifth Edition with African-American students who have been referred for psychoeducational evaluation due to academic and/or behavioral problems.

Significance of the Study

The previous version of the Wechsler scales, Wechsler Intelligence Scale for Children-Fourth Edition, was the most widely used measure of intelligence for over ten years. As of 2013, only six studies had investigated the structure of the WISC-IV in referred samples. The first study applied exploratory factor analysis methods to the WISC-IV core subtest scores of 432 students who were referred for psychoeducational evaluation in Pennsylvania schools (Watkins,

Wilson, Kotz, Carbone, & Babula, 2006). Consistent with the model developed by the publishers, four group factors and a broad general factor were found.

The second study applied confirmatory factor analysis methods to WISC-IV scores from a national sample of 355 students who were referred for evaluation to determine eligibility for special education services (Watkins, 2010). Direct and indirect models fit the data, but the direct hierarchical model was slightly superior. A direct model is one in which each factor is distinguished by the direct effect it has on another without mediation. This is sometimes called a bifactor or nested model. This model provides factor loadings of subtests on a general factor of intelligence (*g*) and indices. An indirect model shows the relationship between higher order factors and a variable as mediated by the lower order factors. In indirect models, such as the model published in the WISC-IV manual, factor loadings demonstrate relationships between *g* and indices and between indices and subtests.

In the third study, Bodin, Pardini, Burns, and Stevens (2009) applied confirmatory factor analysis to a sample of 344 children who received neuropsychological examinations in the Southeastern United States. The fourth study applied confirmatory factor analysis methods to the scores of 550 children with heterogeneous clinical diagnoses who were included in the WISC-IV standardization sample (Chen & Zhu, 2012). In both studies, global fit indices indicated good fit for an indirect hierarchical model, but a direct hierarchical model was not examined.

The fifth study included a sample of 176 Native American students attending three school districts in central Arizona and three school districts in northern Arizona who were evaluated to determine special education eligibility (Nakano & Watkins, 2013). Confirmatory factor analysis was applied. Fit statistics indicated that the ΔBIC favored the indirect hierarchical model, the ΔCFI was neutral, and the ΔRMSEA favored the first-order oblique and indirect hierarchical

models. Given its support by two of the three indices, the indirect hierarchical model was tentatively accepted as the superior fit to this data.

Devena, Gay, and Watkins (2013) employed confirmatory factor analysis to determine the factor structure of the WISC-IV scores of 297 children referred to a children's hospital. Results supported the use of a direct hierarchical model. All studies found that the general factor accounted for the largest proportion of common and total variance. The studies are summarized in Table 1.

Table 1

Investigations of the WISC-IV with Referred Samples

Authors	Participants	Setting	Method	Findings
Watkins, Wilson, Kotz, Carbone, & Babula, 2006	432	Children referred for special education eligibility in Pennsylvania schools	Exploratory factor analysis	Data fit the model developed by the publisher. General ability accounted for 46.7% of total variance and 75.7% of the common variance.
Watkins, 2010	355	National sample	Confirmatory factor analysis	Indirect and direct models fit the data with direct model slightly superior. The general ability factor accounted for 48% of the total variance and 75% of the common variance.
Bodin, Pardini, Burns, and Stevens, 2009	344	Children who received neuro-psychological examinations in southeastern United States	Confirmatory factor analysis	Indirect model like that in the manual. Direct model was not examined.

Chen & Zhu, 2012	550	Subset of standardization sample with heterogeneous clinical diagnoses	Confirmatory factor analysis	Indirect hierarchical model like that in the manual. Direct model was not examined.
Nakano & Watkins, 2013	176 Native Americans	Children referred for special education eligibility in three school systems in northern Arizona and three school districts in central Arizona	Confirmatory factor analysis	An indirect hierarchical model consistent with that in the manual was accepted as superior fit. General ability factor accounted for 33% of the total variance and 69% of the common variance.
Devena, Gay, and Watkins, 2013	297	Patients referred to a children's hospital	Confirmatory factor analysis	Data supported using a direct model. General ability factor accounted for 50% of the total variance and 76% of the common variance.

If this trend continues, studies on the WISC-V that involve referred samples will be rare and most will employ confirmatory factor analysis. This study will be one of the first to examine the validity of the WISC-V with a referred sample of students. In addition, the referred sample will consist of African-American students only. I will examine the data by employing the confirmatory factor analytic method. Measurement invariance between genders will be examined using Mplus 7.4. Furthermore, it will provide insight into the factors that are being measured in the referred sample and assist practitioners in interpreting the scores.

Although the WISC-V manual (Wechsler, 2014) provides evidence of validity for many disability groups, the sample size in each group was very small. Sample sizes ranged from 16 to 95. Each disability group's mean score for each subtest was compared to the mean score for the standardization sample to determine whether there was a significant difference. It is not plausible

to state that the results for such small samples will accurately predict the performance of that subpopulation. This study will include a sample size of 450 students and employ analytic methods that are generally done with large sample sizes (Kline, 2005).

This study will also be significant, because it replicates two other studies. It has long been believed that the route to knowledge is through the accumulation of replicable experimental findings. However, there is a lack of replication studies in the field of psychology. This lack of replications is now considered to be a crisis within the social sciences (Earp & Trafimow (2015).

Assumptions and Limitations

There are several limitations to this study. First, participants were not randomly selected. They consist of African-American students within one school district who have been referred for psychoeducational evaluation because of academic or behavioral difficulties. Therefore, the results may not be representative of the performance of all children within the group that is sampled.

Second, confirmatory factor analysis was used to study measurement invariance. Other approaches, such as item response theory, could provide meaningful, complementary information.

Third, the current sample of students had varying disabilities. It would have been beneficial to have large samples of children with each disability to provide more insight into the effects of race as well as disability on the validity of the WISC-V.

Overview of the Study

This study will examine the factor structure of the WISC-V with a sample of students who have been referred for psychoeducational evaluation to determine eligibility for special education services. Participants will be African-American students in one school district in the

Southeastern United States who are evaluated during the 2015-2017 school years. It is expected that the preferred factor structure for the WISC-V with this sample will be an indirect hierarchical model with general intelligence at the apex (Full Scale IQ), five first-order factors (indexes), and ten second-order factors (subtests). The null hypothesis for this study is the referred sample's data will fit the model derived from the WISC-V normative sample and present no evidence of structural bias. Failure to reject the null hypothesis is evidence that the WISC-V is appropriate for use and interpretation as outlined in the WISC-V *Technical and Interpretive Manual* (Wechsler, 2014) with this subpopulation.

The WISC-V manual offers guidelines for interpreting results. If the model derived from the data collected in this study is configurally different from the model derived from the standardization sample (the data do not fit the published structural model, and we reject the null hypothesis), alternative methods of interpreting WISC-V results may be warranted. For example, it is recommended that practitioners interpret scores by reporting and describing the student's Full Scale IQ, followed by discussions of each factor index score. Significant differences between an index score and the Full Scale IQ represent strengths or weaknesses in those cognitive domains. The clinical importance of such strengths or weaknesses depends upon the child and the context of the evaluation. When significant differences are obtained, corroborating evidence should be provided to support the interpretation of the difference. Additional testing may be warranted to confirm or refute the original hypothesis about the child's profile. If the new data do not support the hypothesis, new hypotheses may emerge. Significance of differences in scores is determined by guidelines in the manual. Those guidelines are based on statistical significance as well as how rare the differences are in the normative sample.

If this study reveals a measurement model that is different, we can conclude that the WISC-V measures different constructs in the standardization sample and the referred sample. Therefore, the published guidelines for interpretation may lead to erroneous inferences being made about students. For example, the WISC-V manual states that for an eleven-year-old child with a Full Scale IQ of 90 to 109, a 10-point difference between his Verbal Comprehension Index and Full Scale IQ is considered statistically significant at the .05 level. A difference of this magnitude occurs in 15% or less of the population, so this difference would be considered statistically significant as well as unusual and worthy of being identified as a strength or weakness for the child. However, if a student in the study obtains a Verbal Comprehension Index score that is over 10 points different from his Full Scale IQ, the difference may not be unusual (as determined by the results of this study) and would not be considered a significant weakness.

2 REVIEW OF THE LITERATURE

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) states that validity is the most fundamental consideration in developing and evaluating tests. Accuracy of an instrument in predicting performance or behavior is directly related to its validity, or degree to which the instrument measures what it is intended to measure. Ary, Jacobs, and Razavieh (1996) defined validity as the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Validation involves gathering evidence to provide an empirical basis for the proposed score interpretations, and validity refers to the degree to which all the accumulated evidence supports those interpretations. The *Standards for Educational and Psychological Testing* specified five sources of validity evidence: evidence based on test content, response processes, testing consequences, internal structure, and relations to other variables.

Cizek, Rosenberg, and Koons (2008) examined information about educational and psychological assessments in the sixteenth edition of the *Mental Measurements Yearbook*. Findings revealed that the total number of validity sources identified was a statistically significant predictor of overall evaluation of an instrument. The more sources of validity evidence reported, the more likely the instrument would receive a favorable evaluation. Each source of validity evidence is discussed below.

Evidence Based on Test Content

Sireci and Faulkner-Bond (2014) conducted a comprehensive review of the literature on evidence of validity based on content and the guidelines outlined in the *Standards for Educational and Psychological Testing*. They concluded that tests cannot be inherently valid or invalid,

because what is to be validated is the use of a test for a particular purpose. Therefore, the first step in validation and test development is to determine the intended uses and interpretations of the test scores. Test developers must consider the extent to which the items on the test correspond to expectations in the content standards, the level of consistency between cognitive complexity articulated in objectives and tested by items, and the extent to which the test reflects the standards in terms of relative emphasis on different topics. In determining an assessment's alignment with content standards, these questions should be answered: "Can everything on the test be found in the state standards? Does each assessment fairly and sufficiently sample the important objectives of the standards? Is each assessment sufficiently challenging (Sireci & Faulkner-Bond, 2014)?"

According to the *Wechsler Intelligence Scale for Children - Fourth Edition* (WISC-IV; Wechsler, 2003) and *Wechsler Intelligence Scale for Children – Fifth Edition* (WISC-V; Wechsler, 2014), evidence of validity based on content is not usually expressed in numerical terms. This evidence is generally based on the degree to which the test items accurately represent and relate to the construct or trait being measured. An investigation of test content also involves evaluating the wording and formatting of items, as well as the procedures for administering and scoring the test. During each research stage of development of the Wechsler instruments, the test developers conducted literature reviews, as well as expert and advisory panel reviews of the items and subtests to improve content coverage and relevance. Experts and advisory panel members were chosen, based on their expertise in child psychology, neuropsychology, and/or learning disabilities. This aspect of validity will not be investigated in this paper, as Pearson provided a detailed description of their process.

Evidence Based on Response Processes

The *Standards* (AERA et al., 2014) described evidence of validity based on response processes as "evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" (p. 15). Recommendations for gathering validity evidence include questioning test takers about their strategies for answering or responses to particular items, maintaining rough drafts of writing tasks, and documenting aspects of performance such as eye movement or response times.

The WISC-IV (Wechsler, 2003) and WISC-V (Wechsler, 2014) manuals indicate that evidence of validity based on response processes should provide support that the child engages the cognitive processes that are expected to be measured by the construct and can be gained from empirical and qualitative examination of response processes. During development of the WISC-IV and WISC-V, response frequencies for multiple choice items were examined to identify commonly given incorrect responses. Respondents were questioned about their incorrect responses and frequently occurring incorrect responses were evaluated for their plausibility as acceptable responses. The subtest, Picture Concepts, required detailed investigation of response processes to ensure scoring reflected the construct being measured. Respondents were asked to explain their reasons for grouping the pictures as they did. Their responses, along with the corresponding response frequencies, provided the rationales for the children's correct and incorrect responses. Based on the observed patterns, items were changed and examined again in future administrations or eliminated from the instrument entirely. A similar procedure was used to examine the new instructions and items for Matrix Reasoning. Direct questioning about the children's understanding of items was utilized to examine response processes on new subtests. Students' answers resulted in additional instruction in the sample item for the Visual Puzzles subtest

and the creation of sample items for the Figure Weights subtest. Other subtests were affected in similar ways by results of children's responses to inquiries or researchers' evaluations of response processes.

Compared to other sources, there are few validation studies aimed at obtaining evidence from response processes (Padilla & Benitez, 2014). Cizek, Rosenberg, and Koons (2008) reviewed validity studies to examine the trends in conducting validation studies. They found that the majority of papers focused on evidence of validity based on content and structure of assessments, and validation studies that focused on response processes comprised only 1.8% of the papers.

Examining this area would require getting the examiners who administer the WISC-V to ask students to explain their problem-solving strategies and reasons for providing certain responses. It is highly unlikely that data would be gathered with fidelity, as this task would require much of the examiners' time, for which, they would not be compensated. Therefore, this source of validity will not be examined in this dissertation.

Evidence Based on Consequences of Testing

Validity evidence based on consequences of testing, often referred to as consequential validity, is highly controversial and contested. It is essentially nonexistent in the professional literature and applied measurement and policy work (Cizek, Bowen, & Church, 2010). Cizek, Rosenberg, and Koons (2008) reviewed validity evidence for the 283 published instruments included in the sixteenth edition of the *Mental Measurements Yearbook*. The instruments were used for several measurement purposes including educational achievement, ability, personality, career guidance, personnel selection, and other areas. They found that evidence of validity based on conse-

quences of testing was noted for only two tests (0.7%), while construct (58.0%), concurrent (50.9%), and content (48.4%) validity evidence were provided fairly frequently. The authors concluded that test producers generally reject evidence of validity related to consequences of testing.

Cizek, Bowen, and Church (2010) sought to determine if findings of Cizek, Rosenberg, and Koons (2008) were idiosyncratic or reproducible. They reviewed articles in eight applied measurement and testing policy-related journals that could potentially publish information related to validity evidence based on consequences. All issues of each journal for the 10-year period between 1999 and 2008 were examined. Of the 2,408 articles published, 1,007 (41.8%) discussed validity, and none of them provided information related to consequences of testing as a source of validity evidence. The researchers also reviewed information presented at the 2007 and 2008 annual meetings of the AERA, APA, and NCME, the three organizations that sponsored the *Standards*. They searched session titles, symposia, individual presentation titles, and keywords for each conference program for the terms: validity, validation, consequences, and consequential. They did not find the intersection of the terms consequential and validity. All presentations that addressed consequences addressed systemic consequences, not those of specific tests or instruments.

Although not commonly researched, there are consequences of testing. Consequences of testing refer to the intended and unintended consequences of *legitimate* test interpretation and use. There are two aspects of these consequences: value implications and personal/social consequences. Value implications include the personal or social values suggested by our interest in the construct and the name selected to represent the construct being measured, personal or social values reflected by the theory underlying the construct, and values reflected by the broader ideo-

logies that impacted the development of the theory. Because the name of a construct or measurement affects the way it is evaluated, one must be careful in naming them. For example, scores on identical instruments titled 'Early Development Instrument' or 'School Readiness Inventory' or 'Developmental Immaturity Scales' will likely be interpreted differently because of the names.

Social consequences of legitimate test use can be positive or negative and intended or unintended. For example, if an instrument is used to screen for or describe depression levels in a population, one needs to consider the consequences of finding very small or very large numbers of depressed people. One should consider how such findings might impact theories about depression and mental health, funding of community mental health programs, and/or group health plan coverage and rates. One must also consider how such findings affect score meaning and use. For example, a state-wide math literacy test might have a potential intended consequence of increased high school graduation rates and a potential unintended consequence of teachers teaching to the test. How do each of the consequences affect the meaning of the state-level math literacy test scores, how 'math literacy' is conceptualized, and theories about math literacy (Hubley & Zumbo, 2011)?

The WISC-IV (Wechsler, 2003) and WISC-V (Wechsler, 2014) manuals state that evaluation of consequences of testing should include unintended consequences of testing, such as item bias and score differences between groups. Information about the consequences of testing may influence decisions to utilize or not utilize a test, but adverse consequences do not detract from the validity of the intended test interpretations.

This area will be indirectly explored, as results of this study will be used to influence the decision to continue the current methods of interpreting results of the Wechsler Intelligence Scale for Children- Fifth Edition or make recommendations to alter the interpretation of assessment results. This study will also be used to determine whether the school district in which the study is conducted will continue use of this instrument.

Evidence Based on Internal Structure

According to the *Standards*, validity evidence based on internal structure refers to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 16). Dimensionality, measurement invariance, and reliability are the three primary aspects of internal structure (Rios & Wells, 2014).

Dimensionality is assessed by determining whether the inter-relationships among the test items support the intended test scores about which inferences will be drawn. Confirmatory factor analysis (CFA) is the most comprehensive means of comparing hypothesized and observed test structures. It examines relationships between indicators and the latent variables (constructs) that the indicators are intended to measure. CFA can evaluate method effects and examine measurement invariance. It is used to verify the number of underlying dimensions and the factor loadings and to provide evidence of how to score an instrument. If a CFA model with one latent variable fits the data well, the instrument should be scored using a composite score. Conversely, a model that has multiple latent variables, and fits the data well, should report each latent variable as a subscale, with factor loadings determining how the subscores should be created. For multi-factor

models, it is possible to detect the convergent and discriminant validity of theoretical constructs (Rios & Wells, 2014).

Measurement invariance is defined as a lack of systematic bias. The main concern in assessing bias is to determine whether knowledge of an examinee's group membership influences the examinee's score on the measured variable, given the examinee's status on the latent variable of interest. One demonstrates test fairness by demonstrating measurement invariance across all distinctive subgroups being evaluated. Standard 3.3 of the 2014 *Standards* states:

"Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 64)."

The last of the three primary aspects of internal structure is reliability. Reliability assesses the reproducibility of test scores on repeated test administrations taken under the same conditions. It is the proportion of true score variance to total observed score variance (Rios & Wells, 2014).

The WISC-IV was often used for clinical assessment, which made it important to demonstrate that WISC-IV scores had the same meaning for children in clinical and non-clinical populations. Because measurement invariance of the WISC-IV across large normative and clinical samples had not been reported, Chen and Zhu (2012) investigated measurement invariance with large samples. They examined whether the WISC-IV core subtests measured latent abilities for the normative sample and clinical samples similarly.

Findings supported the assumption of measurement invariance of the WISC-IV scoring structure between normative and clinical samples. The clinical group scored slightly lower on Comprehension and Coding subtests, which indicated that these subtests measured narrow abilities that were not modeled by the factor structures shown. Children with traumatic brain injury had a significant deficit on the Coding subtest. Lower scores may have been attributed to their lower abilities in processing speed. The ability to make associations and/or fine motor skills may have also accounted for small portions of the deficits. Items on the Comprehension subtest covered some aspects of social functioning. Children with autism and Asperger's disorders, who are known to have major deficits in social functioning, were included in the clinical sample. Thus, explaining lower performance on the Comprehension subtest. The statistically significant differences in performance on the Coding and Comprehension subtests were not large, indicating they were probably not clinically meaningful.

Evidence of validity based on internal structure of the Wechsler Intelligence Scale for Children- Fifth Edition was established through CFA and reported in the WISC-V *Technical and Interpretive Manual* (Wechsler, 2014). This model indicated that a second-order factor, general intelligence (*g*), indirectly influenced subtests through five first-order factors (Verbal Comprehension, Fluid Reasoning, Visual Spatial, Working Memory, and Processing Speed). However, scholars (Canivez, Watkins, & Dombrowski, 2015) have noted that there was insufficient detail in describing how the factors were defined and why weighted least squares estimation was used. Weighted least squares estimation is typically used with categorical or non-normal data, requires much larger sample sizes, and can lead to model misspecification more readily than maximum likelihood estimation. Canivez and colleagues also noted that the preferred CFA model allowed cross-loadings of the Arithmetic subtest, and there was a standardized path coefficient of 1.00

between the general intelligence factor, *g*, and the Fluid Reasoning (FR) factor, which suggested that *g* and FR were empirically redundant. Canivez et al. also expressed concern about use of chi-square difference tests of nested models to identify the five-factor model, because this approach is known to be misleading when the base model is misspecified and is overly powerful with large sample sizes (Dombrowski, Canivez, Watkins, & Beaujean, 2015).

There are five additional areas of concern with NCS Pearson, Inc.'s approach to documenting the structure of the WISC-V. First, rival models, such as a bifactor model, were not examined. Bifactor models are sometimes preferred over higher-order models for tests of cognitive ability, because they allow for partitioning of general and group factor variance and are considered to be more comparable with Carroll's three-stratum theory of cognitive ability. Rival models would aid clinicians and researchers in determining the interpretability of group factors (Dombrowski, Canivez, Watkins, & Beaujean, 2015).

Second, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) emphasized the need for statistics such as model-based reliability estimates including omega-hierarchical and omega-subscale in IQ test manuals that recommend the interpretation of subscores. Omega estimates can aid in determining how much interpretive emphasis should be placed upon scores designed to measure primary and secondary factors. Model-based reliabilities can be estimated with Omega-hierarchical and Omega-hierarchical subscale. Omega-hierarchical is the reliability estimate for the general intelligence factor with variability of group factors removed. The reliability estimate of each group factor, Omega-hierarchical subscale, can be calculated with all other groups and general factors removed. Omega values can be produced with the *Omega* program (Watkins, 2013). Omega coefficients should exceed .50, but .75 is pre-

ferred (Canivez, Watkins, & Dombrowski, 2017). These statistics were not included in the WISC-V *Technical and Interpretive Manual*.

Third, NCS Pearson, Inc. chose to rely upon CFA procedures when providing evidence of validity based on internal structure and did not provide results of exploratory factor analyses (EFA). Reise (2012) indicated that EFA and CFA are complementary and test users are more confident about the instrument's structure when both models are in agreement, especially when an instrument has been revised and reformulated. For example, elimination of some subtests found on the WISC-IV and addition of new ones on the WISC-V could have caused unexpected changes to the WISC-V factor structure that would benefit from EFA prior to use of CFA (Dombrowski, Canivez, Watkins, & Beaujean, 2015). Fourth, independent research on intelligence test factor structures using EFA methods have produced different results from those using CFA models of existing subtests (Dombrowski, Canivez, Watkins, & Beaujean, 2015).

Finally, Canivez et al. (2015) employed EFA methods using the Schmid-Leiman (SL) procedure to investigate the WISC-V total sample correlation matrix. This procedure mathematically transforms a second-order structure into an orthogonal first-order factor structure, where general (*g*) and group (Verbal Comprehension, Fluid Reasoning, Visual Spatial, Working Memory, and Processing Speed) factors both directly influence indicator variables. It was argued by its developers that this process preserves the desired characteristics of the oblique solution and discloses the hierarchical structure of the variables. Canivez et al.'s SL analysis resulted in a four-factor solution that mirrored the WISC-IV's factors. Fluid reasoning and visual spatial combined to form the WISC-IV's previously identified perceptual reasoning factor. The four first-order factors accounted for small portions of the total and common variance, while the second-order *g* factor accounted for large portions of the total and common variance. It was recom-

mended that clinical interpretations of the WISC-V should be primarily, if not exclusively, at the general intelligence level.

Loading values on higher-order models may be biased, if there are cross-loadings and loadings of all measured variables on a group factor are constrained to be proportional. Therefore, an alternative to the SL procedure for EFA was developed by Jennrich and Bentler in 2011: exploratory bifactor analysis (EBFA). EBFA posits that the general factor of intelligence directly influences performance on subtests instead of indirectly influencing performance through a first-order or group factor. It was thought to be better than the SL transformation (Dombrowski, Canivez, Watkins, & Beaujean, 2015). However, the only independently published article comparing SL procedure for EFA and EBFA on cognitive ability data found consistent results for the two procedures (Dombrowski, 2014).

Dombrowski et al. (2015) examined the WISC-V factor structure using exploratory bifactor analysis. Data consisted of the WISC-V subtest correlation matrix for the total standardization sample that was published in the *Technical and Interpretive Manual* (Table 5.1; Wechsler, 2014). Results indicated that a three-factor model, instead of the five-factor model suggested by the test developers or the four-factor model derived by employing SL transformation, was the most plausible for the WISC-V. The three factors were: Processing Speed, Working Memory, and Perceptual Reasoning. Block Design, Visual Puzzles, Matrix Reasoning, and Figure Weights converged to form a single factor, Perceptual Reasoning, as opposed to the separate factors, Visual Spatial and Fluid Reasoning that were suggested by the publishers of the WISC-V. None of the models showed evidence of definitive Fluid Reasoning or Verbal Comprehension factors. Lack of a Verbal Comprehension factor is inconsistent with the structure presented in the *Technical and Interpretive Manual* and with existing literature on the Wechsler scales.

Some reasons for the differences in findings using the SL transformation and exploratory bifactor model were offered. In bifactor models, all factors, including *g*, are first-order factors. All factors compete to explain the subtests' covariance. Typically, *g* is formed first, and group factors are formed from the remaining covariance unexplained by *g*. In higher-order models, first-order factors are formed first, then, *g* is formed from the first-order oblique factors. The SL rotation of the higher-order models simultaneously calculates all the subtests' indirect relationships to *g* and group factors' residuals/error. For *g* to have strong indirect relationships to subtests, subtests must have strong loadings on group factors and group factors must have strong loadings on *g*. It appears that the Verbal Comprehension factor was not strongly defined or the verbal subtests had strong loadings on a different factor, so *g* did not have a sizeable relationship with verbal subtests.

In the exploratory bifactor analysis, *g* explained all that is common among the verbal subtests, so the residual covariance from these tests was attributed to specific factors and error variance. It is not unusual to find some group factors diminish when using a bifactor model/rotation, if the general factor is well defined.

Dombrowski et al.'s (2015) results revealed that the WISC-V is primarily a measure of *g*, because it accounts for a majority of the subtests' total and common variance. These findings are consistent with other studies of Wechsler scales using both EFA and CFA methods (Bodin, Pardini, Burns, & Stevens, 2009; Watkins, 2010). Again, it was recommended that interpretation of the WISC-V should focus on the general factor.

In this dissertation, I will examine dimensionality and measurement invariance, but reliability will not be addressed. Research has been conducted on the WISC-V standardization sam-

ple using confirmatory factor analysis (Wechsler, 2014), exploratory factor analysis (Canivez et al., 2014, 2015), and exploratory bifactor analysis (Dombrowski, 2015). Factor structure of the previous version of the Wechsler scale (WISC-IV) was investigated with a referred sample six times. Five of the six studies used CFA (Watkins, 2010; Bodin, Pardini, Burns, & Stevens, 2009; Chen & Zhu, 2012; Nakano & Watkins, 2013; Devena, Gay, & Watkins, 2013), and one used EFA (Watkins, Wilson, Kotz, Carbone, & Babula, 2006).

Initially, my intentions were to conduct an exploratory factor analysis of data obtained from a referred sample of African-American students. However, it was decided that a study that employs confirmatory factor analysis, a method that has been used repeatedly with the standardization samples and referred samples on different versions of the Wechsler scales, would be more beneficial in allocating results of the study to the current sample as opposed to the research methods. Therefore, confirmatory factor analyses will be employed to determine whether the WISC-V is measuring the same constructs in the referred sample as it measures in the standardization sample.

Measurement invariance can be assessed in numerous ways; however, the focus for this dissertation will be on multiple-group confirmatory factor analysis (MGCFA) or factorial invariance, which assesses invariance at the scale-level. It allows for simultaneous model fitting across multiple groups, assesses various levels of measurement invariance, controls for measurement error, and utilizes direct statistical tests to evaluate cross-group differences of the estimated parameters.

There are three aspects of factorial invariance- configural invariance, measurement invariance, and structural invariance. Configural invariance, the most basic and necessary condition

for comparing groups, is reflected when identical indicators measure the same latent construct(s) of interest across groups. Measurement invariance refers to (a) metric invariance – equal factor loadings across groups, (b) scalar invariance – equal item intercepts across groups, and (c) invariance of item uniqueness – equal item error variances/covariances across groups (Dimitrov, 2010). Measurement invariance or metric equivalence, a more restrictive form of invariance, assumes configural invariance as well as equivalent strengths (factor loadings) between indicators (items) and latent constructs across groups (Rios & Wells, 2014). Attainment of metric equivalence denotes equal measurement units of the scale designed to measure the same latent constructs across groups. Therefore, the relations between the latent construct and external variables can be compared across groups, because a one-unit change in one group would be equal to a one-unit change in any other group. However, the construct means cannot be compared, as the origins of the scale may differ across groups (Dimitrov, 2010). To make direct comparisons of latent group means, scalar equivalence is necessary.

Attainment of scalar equivalence requires configural and metric equivalence, as well as equal item intercepts across groups. When one is interested in the equivalence of covariances among groups for a number of latent constructs within the model, scalar equivalence may be required (Rios & Wells, 2014). The lack of invariant intercepts indicates item bias or differential item functioning. For example, students from different groups that have equal verbal ability may perform differently on some items in a verbal ability test because of offensive language in these items that offend the students in a particular group.

Even more stringent is the invariance of item uniqueness. This level requires equal factor loadings, equal indicator intercepts, and equal item uniqueness across groups. Invariance of item uniqueness across groups suggests that the items were measured with the same precision in each

group and group differences on any item are due only to group differences on the common factors (Dimitrov, 2010).

Structural invariance refers to invariance of factor variances and covariances. Testing for structural invariance is done when the variability of target constructs and/or correlational relationships among them are considered relevant to the generalizability aspect of validity. When evaluating factorial invariance, the aspects to test would depend on the specific validation goals. For example, equal factor loadings and equal intercepts across groups (strong measurement invariance) are needed to compare groups on factor means, while construct validation requires weak measurement invariance in testing for equivalence of factor variances and/or covariances across groups (equal factor loadings) only (Dimitrov, 2010). In this study, the ultimate goal is to examine structural invariance across groups. However, group comparisons on factor variances and/or covariances are meaningful only when the factor loadings are invariant. Therefore, configural, measurement, and structural invariance will be investigated.

Measurement Invariance between African-American Females and Males

When examining an instrument's invariance, researchers often investigate invariance between genders. This seems logical, because the males and females within the population being studied have usually lived within the same environment and have been exposed to similar stimuli. For example, they usually have gone to the same schools and churches and have participated in or observed the same activities. It can be argued that performance differences between the groups would indicate measurement bias. However, in some cultures, males and females are socialized differently. When invariance is upheld in those populations, it is further evidence that the instrument is psychometrically sound.

In the African-American community, there is a popular saying, "Mothers love their sons, but raise their daughters." This phrase suggests that African-American mothers give their sons warmth and affection, while imposing higher expectations, responsibility, and restrictions on their daughters. Thus, they are better preparing their daughters than their sons for their future adult roles. Several researchers have argued that this differential in socialization has resulted in large discrepancies in academic achievement between African-American adolescent males and females (Varner & Mandara, 2013). African-American females have higher middle and high school grade point averages, standardized test scores, high school graduation rates, college matriculation and graduation rates, academic self-efficacy, school importance beliefs, and academic expectations than African-American males. While differences in academic achievement are apparent, those differences cannot be definitively attributed to differential parenting practices. Varner and Mandara (2013) examined differential parenting of African-American children based on gender and birth order as an explanation for achievement differences between genders. Results of the study revealed girls reported receiving more monitoring, communication, and rule enforcement, but less autonomy to make decisions than their younger male siblings. Mothers reported higher expectations for girls than for boys. Mothers who were more concerned that their children would experience racial discrimination had lower academic and behavioral expectations for their children who they thought were most at risk for discrimination, usually the males, so they parented them differently. A significant percentage of the grade point average and test score gap was accounted for by parenting differences. It was suggested that reducing differential parenting could help alleviate differences in achievement among African-American students.

In another study, Mandara, Varner, and Richman (2010) found that parenting based on gender and birth order contributed to academic differences in African-American students. First-

born girls had the highest achievement scores. Later-born boys, who lived in less cognitively stimulating homes, were given fewer household chores, who argued more with their mothers, and who had less latitude in decision-making, had the lowest achievement scores. There was no significant difference in the achievement of first-born boys and later-born girls. It was suggested that the later-born African-American boys would achieve at the same rates as their siblings, if they were socialized in the same manner as their siblings.

In addition to parenting, partial reasons for the differences in academic achievement might be a decline in academic motivation. Researchers believe that African-American boys and girls begin school equally motivated to learn, but boys differentially lose this motivation (Ogbu, 2003; Osborne, 1995). McMillian, Frierson, and Campbell (2011) attempted to answer the question: When do gender differences emerge?

McMillian, Frierson, and Campbell (2011) investigated whether a gender gap exists in self-rated academic competence and global self-esteem measures in middle childhood. They examined gender differences in academic identification, which is based on students' desires to do well in school, with positive self-esteem linked to success in school. Historical data were reviewed on 113 African-American students enrolled in predominantly White public schools in the southeastern United States. Differences in students' self-esteem, academic self-concept, and academic accomplishment at ages 8 and 12 were compared.

Academic performance is expected to be a major source of global self-esteem and academic self-concept for students. Global self-esteem has been defined as the assessment of personal worth that people place on themselves, based on what they think they have accomplished, as well as what they perceive to be others' opinions of them. Following this logic, elementary

school-aged African-American boys might have lower self-esteem than African-American girls, because they tend to perform less well in academics. However, research has shown that, across racial groups, boys tend to have slightly higher self-esteem than girls. Also, a meta-analysis of 261 studies involving more than 500,000 cases of children, adolescents, and young adults revealed higher self-esteem in African-Americans than in Whites at all age levels. This is unexpected, given the continuing achievement gap between African-Americans and Whites. It implies that self-rated competence may contribute less to overall self-esteem for African-Americans than for Whites (McMillian, Frierson, & Campbell, 2011).

Academic achievement is more predictive of academic self-concept than general self-esteem, and it predicts academic self-concept to a greater degree in African-American girls than in African-American boys. The relationship between academic self-concept and academic achievement decreases in male college students, but does not decrease in female college students. Given the fact that African-American girls outperform African-American boys academically, yet, the boys maintain equal levels of self-esteem, it has been hypothesized that African-American males' self-esteem does not depend on academic achievement (Cokley & Moore, 2007). McMillian, Frierson, and Campbell (2011) found no gender differences in reading or mathematics achievement between African-American boys and African-American girls at ages 8 or 12. Self-esteem was predicted by academic performance in similar ways for both genders.

In 2016, McMillian, Carr, Hodnett, and Campbell examined whether there were gender differences among these same students at age 15. These were the hypotheses: Fifteen-year-old girls would outperform 15-year-old boys in mathematics and reading. Previous achievement in mathematics and reading would be less predictive of academic self-concept for 15-year-old African-American boys compared to African-American girls. The relationship between academic

self-concept and global self-esteem will be stronger for 15-year-old African-American girls than for 15-year-old African-American boys. Achievement in mathematics and reading would be less related to global self-esteem for African-American boys than for African-American girls at age 15. No gender differences were found in academic achievement or the extent to which academic achievement diminishes a student's perception of academic competence. There was no statistically significant relationship between academic achievement and global self-esteem for 15-year-old African-American boys and girls. However, it appeared that African-American girls valued academics more than their male counterparts.

Skinner, Perkins, Wood, and Kurtz-Costes (2016) summarized literature on gender development in African-Americans and described recent findings on socializing factors, particularly parenting and media, which may shape that development. The literature suggested that African-American boys feel more pressure for gender conformity than their female peers. During early childhood, boys' and girls' occupational preferences are consistent with traditional gender roles. However, by adolescence, African-American girls may be more likely to aspire to professional careers. African-American children endorse traditional gender stereotypes with girls having stronger verbal skills and boys having more athletic ability. Females are more likely to report competence in masculine domains than males are to report competence in traditionally feminine domains. In early and middle childhood, African-American females tend to prefer same-sex, same-race peers, while males prefer male peers, regardless of race. By adolescence, African-American girls have better quality same-sex friendships than boys. They also have more egalitarian gender attitudes than their male peers.

Gender socialization within the family and through media facilitates much of the gender development discussed above. Again, research has shown that African-American mothers treat

their sons and daughters differently (Varner & Mandara, 2013). Girls typically receive more be-havioral control than boys. Boys are allowed more autonomy than girls in areas such as staying home alone, attending parties, dating, and sexual activity. Mothers also monitor their daughters' friendships more than they monitor their sons' friendships (Varner & Mandara, 2013). In con-trast, fathers of preschool-aged boys are more involved in their children's activities than fathers of pre-school aged girls (Skinner et al., 2016).

Racial socialization, the process through which parents teach their children about the meaning of being Black in the United States, is an important area of socialization. Theorists be-lieve that women are responsible for passing on cultural values and males are more likely to ex-perience racial discrimination than females. Therefore, parents talk to their daughters about cul-tural pride and their sons about discrimination. African-American mothers perceive girls as aca-demically stronger in all areas than boys and expect more from their girls academically. Howev-er, those mothers also adhere to the stereotypes that boys are better at math than English, while girls are better at English than math. These beliefs often feed into the causal attributions of their children's academic outcomes. Mothers who believe that boys are better at math tend to attribute their sons' math achievement to high math ability, and their sons have higher perceptions of their math competence than boys whose mothers do not believe the gender math stereotypes. This shows that mothers' beliefs about gender differences may shape their children's beliefs, with possible impacts on their developing perceptions of competence and identity (Skinner et al., 2016).

African-American youth have high rates of media consumption. Reports using a national-ly representative sample of 8- to 18-year-olds stated that African-American youth consume al-most 13 hours of media per day. Media exposure and identification with media characters in tel-

evision and videos are related to African-American adolescents' beliefs about attributes that are ideal for women and men, gender attitudes, and self-perception. In a study, those who watched more music videos favored more traditional gender role attitudes and sexual stereotypes than those who had lower rates of media consumption (Skinner et al., 2016).

All of these factors, parenting, motivation, academic self-concept, self-esteem, gender development, racial socialization, and media consumption, affect students' performance. African-American boys and girls are parented and socialized differently, which manifests in different perceptions about academic performance and motivation. These factors may also affect students' perceptions about intelligence tests and motivation to perform on those tests.

Measurement variance between genders can be attributed to one or more of these factors, as well as other factors that were not discussed. Its presence would suggest that gender is a predictor of one's performance on this instrument and necessitate further research to investigate the reason for the finding. On the contrary, providing evidence of measurement invariance would indicate that one's gender has no effect on one's performance on the WISC-V. This will add to the literature supporting the publisher's claim of psychometric soundness.

Evidence Based on Relations with Other Variables

The validity source, relations to other variables, is often expressed in terms of how accurately test scores predict criterion performance or how well an assessment predicts an individual's performance on a specified criterion (Glover & Albers, 2007). The criterion variable is determined by the test users. University grades, school grades, or achievement test scores are variables in relation to which the validity of an instrument can be evaluated. High-stakes tests used in an economically and culturally diverse population need strong bodies of empirical evidence to

justify their use (Oren et al., 2014). There are two types of criterion-related validity: predictive validity and concurrent validity.

The WISC-IV (Wechsler, 2003) and WISC-V (Wechsler, 2014) manuals discussed the relations of those instruments with two types of external variables: scores on other instruments created to measure the same or similar constructs and subjects' designation as members of special groups (e.g., intellectually gifted, intellectual disability, learning disability, autism, etc.). Relations between the WISC-V and the following external instruments were examined in nonclinical samples: Wechsler Intelligence Scale for Children- Fourth Edition, Wechsler Preschool and Primary Scale of Intelligence- Fourth Edition, Wechsler Adult Intelligence Scale- Fourth Edition, Kaufman Assessment Battery for Children- Second Edition, Kaufman Test of Educational Achievement- Third Edition, Wechsler Individual Achievement Test- Third Edition, Vineland Adaptive Behavior Scale- Second Edition, and Behavior Assessment System for Children- Second Edition Parent Rating Scales.

One of the greatest concerns in research has been whether cognitive ability tests represent equivalent assessments for each racial/ethnic subgroup and whether test scores relate to performance criteria equally for each subgroup (Berry, Clark, & McClure, 2011). This issue has been investigated through methods of determining differential validity and differential prediction.

Berry, Clark, and McClure (2011) quantitatively summarized existing differential validity evidence by meta-analyzing the criterion-related validity of a broad range of cognitive ability tests used for selection and placement purposes for Asians, Blacks, Hispanics, and Whites- the four racial/ethnic groups for which differential validity research is most widespread. Their study represented the largest examination to date of racial/ethnic differential validity evidence for cog-

nitive ability tests. Studies including over one million participants across and within educational admissions, civilian employment, and military literatures were examined in the meta-analysis.

Differential validity focuses on the differences between correlation coefficients of tests such as cognitive ability tests (predictors) and measures of performance such as academic achievement (criteria) across subgroups. Differential prediction focuses on differences between unstandardized regression slopes and intercepts relating test and criterion across subgroups. It is the preferred method of comparing predictor- criterion relationships, as it most directly addresses whether test scores predict equivalent criterion scores across subgroups. The unstandardized regression coefficient is not affected by direct range restriction on the predictor, as is the correlation coefficient. Differential prediction can also include separate comparisons of slopes and intercepts, making the method more informative than the correlation coefficient, which contains no information about the differences in intercepts. However, an examination of differential validity as well as of differential prediction evidence has the potential to provide more information than an examination of differential prediction alone.

There are at least four categories of factors that present evidence of differential validity among subgroups: range restriction, contextual influences, psychometric characteristics of tests or criteria, or true differences among subgroups in the role that the predictors play on the criteria. Each is explained below.

Sometimes, true validity of a test may not differ by subgroup, but greater amounts of range restriction in test scores of subgroups can cause observed validity to be lower for those groups that exhibit a smaller range of scores. The differences between subgroups in range restriction influences variance, which alters validity estimates.

Contextual influences can affect the validity of test scores between subgroups, if their affects differ systematically for the groups. For example, if minority test takers feel excessive stress caused by fear of confirming a negative stereotype about their group, they may perform less well during an assessment. The variance due to stereotype threat would act as construct-irrelevant variance that could cause test scores of minority test takers to be less related to the true criterion being assessed.

Psychometric characteristics of tests, such as measurement error and measurement bias, can differ between groups and cause differences in validity. Measurement error is usually assessed through reliability estimates. Measurement bias is typically assessed through differential item functioning and measurement invariance. Differential item functioning assesses measurement bias associated with individual items on tests/criteria, while measurement invariance investigates measurement bias at the test or scale level, typically through factor structures.

In their meta-analysis, Berry, Clark, and McClure (2011) operationalized cognitive ability test measurement invariance as the degree to which the factor structure of a cognitive ability test is equivalent for minority and majority groups (factorial invariance). If the factor structure is different for subgroups, the psychological meaning of test scores is different for the subgroups and could affect the degree to which test scores are predictive of performance criteria. They found that research studies to date do not support the idea of widespread internal psychometric characteristics of tests varying between subgroups, but there is some evidence of differences in prediction of performance criteria.

Across the three broad fields (educational admissions, military, and civilian employment) that commonly use cognitive ability tests for high-stakes selection and placement, evidence sup-

porting lower criterion-related validity for Black samples than White samples was relatively common. This was true for Hispanics as well, although data were only available in the educational admissions setting. The Hispanic-White observed validity difference was smaller than the Black-White observed validity difference. Asian-White validity data were only available in the educational settings, and the validity difference was small to nonexistent. Berry, Clark, and McClure (2011) demonstrated that the evidence is supportive of differential validity and highlighted the need for future research investigating causal factors.

The sizes of the validity differences in the meta-analysis were appreciable, although at first glance, they may seem small. Validity was .04 higher for Whites than for Blacks and Hispanics in educational admissions. When viewed in terms of test utility and percentages, Black and Hispanic validity was 11.8% ([.34-.30]/.34 = .118) lower than White validity in the educational admissions domain. Test utility is a function of the validity of the test. A reduction of 11.8% in the validity of a test means that, holding all other factors constant, the utility of the test is 11.8% lower for Blacks and Hispanics than for Whites. This can cause substantial differences between the groups in the rates of false negative and false positive, resulting in many qualified Black and Hispanic students being denied admission and many unqualified White students being granted admission to educational institutions (Berry, Clark, & McClure, 2011).

Evidence of predictive bias can have serious implications for school systems that utilize data gathered from those instruments to make decisions about interventions. Differential prediction is also an important aspect of consequential validity, because potential bias in predictive scores can differentially effect decisions for different groups, and thus introduce construct irrelevant variance into the decision-making process (Betts, 2008).

Although five types of validity evidence are described in the *Standards*, three types of validity evidence are normally used in evaluating instruments: criterion-related, structural, and content-related. For purposes of this study, only evidence based on internal structure will be examined.

Interpretation of the Wechsler Intelligence Scale for Children- Fifth Edition Scores

As stated in the introduction, validity is about the meaning of scores and the degree to which our inferences are appropriate. What are validated are the inferences, interpretations, actions, or decisions that are made based on test scores.

The WISC-V *Technical and Interpretive Manual* (Wechsler, 2014) recommends that the practitioner evaluate results within the context of the referral question or purpose of the evaluation. Cognitive scores should be interpreted in conjunction with a thorough personal history and clinical observations. According to the WISC-V manual, when interpreting performance on the WISC-V, practitioners should begin by reporting and describing the Full Scale IQ score. Next, each of the factor index scores is reported and described. The third step involves comparisons of the primary (factor) index scores to the Full Scale IQ. A significant difference between the two scores represents strength or weakness in that cognitive area. Next, factor index level pairwise comparisons are made, followed by evaluation of subtest-level strengths and weaknesses. Finally, subtest-level pairwise comparisons are made.

Based on the findings of other scholars (Dombrowski, Canivez, Watkins, & Beaujean, 2015; Canivez, Watkins, & Dombrowski, 2015, 2017), whose research supported only interpreting the Full Scale IQ, the methods of interpretation that are recommended in the WISC-V manual cannot be validated and may lead to erroneous decisions with regard to student placement. An

example of this would be a student who is classified as having a specific learning disability, after

a practitioner uses score differences on the WISC-V to identify processing deficits or weakness-

es. This is a very common practice, and I am surprised to know that research directly contradicts

this practice.

3 METHODOLOGY

The purpose of this study is to investigate the validity of the Wechsler Intelligence Scale for Children- Fifth Edition with African-American students who have been referred for psychoeducational evaluation due to academic and/or behavioral problems. There are five sources of validity evidence described in the *Standards* and the WISC-V manual; however, only evidence of validity based on internal structure will be examined in this study.

Evidence of validity based on internal structure is most often evaluated through factor analysis. The purpose of factor analysis is to identify a set of factors and the structure of their relationships with one another. Specific relations between observed variables (subtest scores) and latent variables (constructs such as working memory), and among latent variables are specified in the model. It is a tool for identifying the fewest factors that account for the pattern of the data. One of the most important outcomes of factor analysis is an understanding of the number and nature of the factors necessary to explain how subtests interconnect. Exploratory factor analysis is intended to help generate new theories by applying statistical algorithms to data to develop latent factors that best account for the variations and interrelationships of measured variables. Once an optimal factor model is identified, it can be evaluated with a type of structural equation modeling known as confirmatory factor analysis. Confirmatory factor analysis is different from exploratory factor analysis. Confirmatory factor analysis is used to test the theories that are developed through exploratory factor analysis. The pre-specified models are tested to determine if they provide reasonably good, yet parsimonious, explanations of the correlations among subtests (Wechsler, 2014).

Conceptual Framework

Validity based on internal structure was assessed by determining whether the WISC-V measures the same factors in the referred sample as it does in the standardization sample. This was done through confirmatory factor analysis by fitting the data from this sample to the model provided in the WISC-V manual that was derived from the scores of the standardization sample. Measurement invariance between genders was determined by comparing the models for boys and girls in the referred sample.

Participants

Participants were 607 children between the ages of 6 years, 0 months and 16 years, 11 months who were referred for psychoeducational evaluation and assessed with the Wechsler Intelligence Scale for Children- Fifth Edition during the 2015-2017 school years, by one of 30 school psychologists who work for an urban school district in the Southeastern United States.

The total sample of 607 students included 531 African-Americans, 42 Whites, 12 Hispanics, 2 biracial students, 1 Native American, and 19 racially unidentified students. The racially unidentified students withdrew from the school system before the study was conducted, so their records were no longer available for review to determine their racial identities.

The 531 records of African-American students were examined. Records which included all 10 primary subtest scores, 5 index scores, and the Full Scale IQ score were retained. Many records did not have the 10 primary subtest scores, because examiners have the option of substituting a supplemental subtest for a primary subtest when administration of a primary subtest has been spoiled. Also, sometimes, students have disabilities that will prevent them from completing all of the primary subtests. For example, the Block Design subtest will not be administered to students with limited use of their hands, so a supplemental subtest will be administered or all

subtests that comprise that index will be omitted. This resulted in 478 cases having sufficient data to be included in the study. Five of those cases were identified as outliers by Mahalanobis distance statistics, which left 473 cases to be included in the study.

Males comprised 62 percent (n = 295) of the sample and females comprised 38 percent (n = 178) of the sample. Participants in the study represented heterogeneous disabilities. Thirty-eight percent (n = 180) were served in the special education program for students with Specific Learning Disabilities, 9 percent (n = 44) were considered Other Health Impaired, 8 percent (n = 38) had Mild Intellectual Disabilities, 5 percent (n = 25) had Emotional/Behavioral Disorders, 3 percent (n = 13) had Speech/Language Impairments, 2 percent (n = 11) were diagnosed with Autism Spectrum Disorder, 2 percent (n = 8) had Significant Developmental Delays, 0.4 percent (n = 2) had Moderate Intellectual Disabilities, and 0.2 percent (n = 1) was considered Deaf/Hard of Hearing. Two percent (n = 8) had been evaluated, but the meeting to determine eligibility for special education services had not been held, so their disability areas, if any, had not been determined. Fifteen percent (n = 73) of the students were determined not to have disabilities, and 15 percent (n = 70) of the students' records were removed from the computer system due to graduation or transfer, so their areas of disabilities, if they had any, could not be determined.

The students ranged in grade from kindergarten to tenth grade. English was the primary language spoken by all participants. Some of the evaluations were the students' initial evaluations, which were conducted to assist in determining whether the students qualified for special education services. Some were reevaluations, which were conducted to determine whether the students who were already receiving special education services continued to qualify for those services.

The publishers used the scores of 2200 children to standardize the WISC-V. The standardization sample was divided into 11 age groups with 200 children in each group. The normative sample was stratified in each age group to match the October 2012 United States Census Bureau for demographics such as ethnicity, parental education, and geographic region. An equal number of males and females were included in each age group. Data for all age groups were used to derive one structural model, which was published in *the WISC-V Technical and Interpretive Manual*.

Instruments

The Wechsler Intelligence Scale for Children- Fifth Edition is an individually administered intelligence test designed for children age 6 years, 0 months to 16 years, 11 months. It is composed of 10 primary subtests, six secondary subtests, and five complementary subtests. The primary subtests are Similarities (SI), Vocabulary (VC), Block Design (BD), Visual Puzzles (VP), Matrix Reasoning (MR), Figure Weights (FW), Digit Span (DS), Picture Span (PS), Coding (CD), and Symbol Search (SS), while the secondary subtests are Information (IN), Comprehension (CO), Picture Concepts (PC), Arithmetic (AR), Letter-Number Sequencing (LN), and Cancellation (CA). The five complementary subtests are Delayed Symbol Translation, Immediate Symbol Translation, Naming Speed Literacy, Naming Speed Quantity, and Recognition Symbol Translation.

The WISC-V is organized into four levels. The Full Scale IQ is comprised of seven primary subtests across five factors, Verbal Comprehension, Visual- Spatial, Fluid Reasoning, Working Memory, and Processing Speed. If one of the Full Scale IQ subtests is invalid or missing, that subtest can be substituted by a secondary subtest from the same factor. Only one substitution is allowed. The primary scale level is composed of the 10 primary subtests that are used to

estimate the five factor index scores. No substitutions are allowed for the Primary Index Scales. The Ancillary Index level is made of five scales that are not included in the factor structure: Quantitative Reasoning, Auditory Working Memory, Nonverbal, General Ability, and Cognitive Proficiency. These scales are composed of various combinations of primary and secondary subtests. The Complementary Index level includes three scales: Naming Speed, Symbol Translation, and Storage and Retrieval derived from the complementary subtests. Complementary subtests cannot be substituted for primary or secondary subtests. The average internal consistency reliability estimates range from 0.88 to 0.96 for composites, 0.81 to 0.94 for primary subtests, and 0.82 to 0.90 for secondary subtests. Although there is no definite standard for determining the quality of reliability coefficients, generally, reliability coefficients around 0.90 are considered excellent, values around 0.80 are considered very good, those around 0.70 are considered adequate (Kline, 2005, p. 59).

Procedures

The development of the WISC-V was based on the assumption that the instrument provides an estimate of general cognitive ability that manifests itself in five cognitive domains – Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed. This is considered a second-order factor model with five first-order factors (cognitive domains) and one second-order factor (general intelligence or *g*).

Evaluating Measurement Models

Mplus version 7.4 was used to conduct confirmatory factor analyses using maximum likelihood estimation. Covariance matrices were produced for the analyses using the correlation matrix, means, and standard deviations from the referred sample.

Confirmatory factor analyses were performed by replicating the procedures utilized by Canivez, Watkins, and Dombrowski (2017), when they examined the factor structure of the 10 primary subtests of the WISC-V standardization sample. That five-factor hierarchical model for the primary subtests was published on page 84 of the WISC-V *Technical and Interpretive Manual* and presented in Figure 1. Goodness-of-Fit indices were:

$$\chi^2 = 135.5, df = 25, \text{CFI} = 0.99, \text{TLI} = 0.98, \text{RMSEA} = 0.05.$$

The structural models presented in Table 5.3 of the WISC-V *Technical and Interpretive Manual* that pertain to the primary subtests were reproduced as indicated in Table 3.
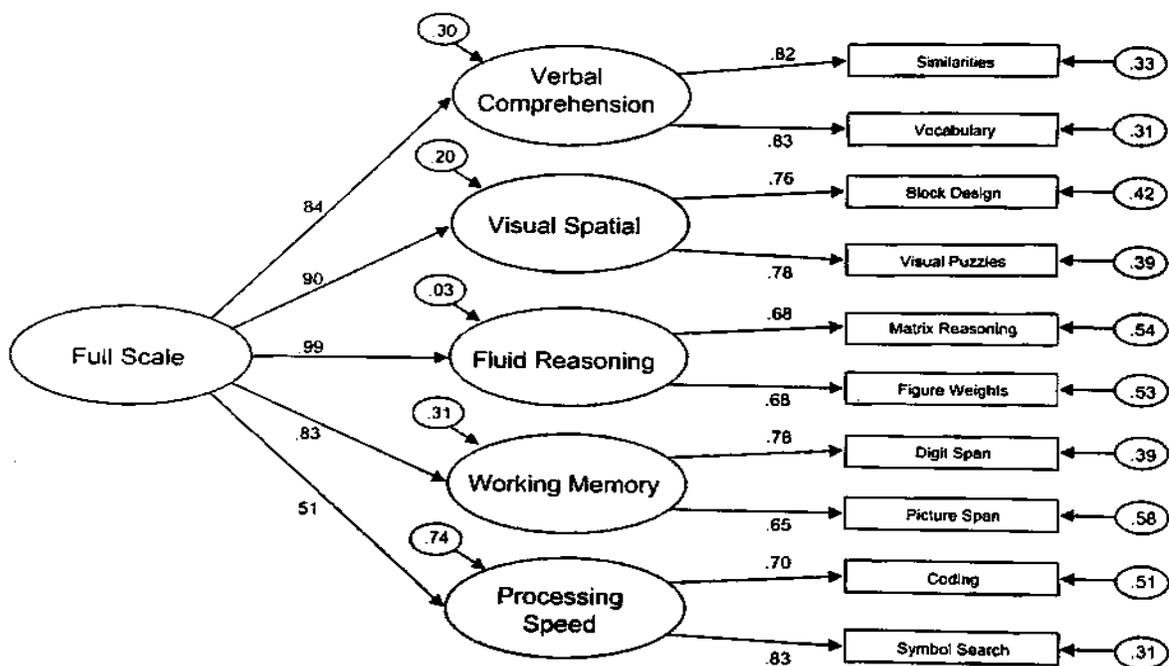
Figure 1. Five Factor Hierarchical Model for the Primary Subtests.

Overall model fit was evaluated using the comparative fit index (CFI), standardized root mean squared residual (SRMR), Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). With the CFI and TLI, higher values indicated better fit, while with

the SRMR and RMSEA, lower values indicated better fit. Criteria for adequate model fit were CFI and TLI ≥ 0.90, SRMR ≤ 0.09, and RMSEA ≤ 0.08. Criteria for good model fit were CFI and TLI ≥ 0.95 with SRMR and RMSEA ≤ 0.06. Models were considered superior, if they exhibited adequate to good overall fit and displayed meaningfully better fit than alternative models (ΔCFI > 0.01 and ΔRMSEA > 0.015). Akaike Information Criterion (AIC) was considered. AIC does not have a meaningful scale, so the model with the smallest AIC was preferred, because it is the most likely to replicate (Canivez, Watkins, & Dombrowski, 2015).

Evaluating Measurement Invariance

Tests of invariance were based on the analysis of covariance structure models using Mplus version 7.4. First, normality of each subtest was evaluated. Maximum likelihood estimation is known for its robustness and is adequate for data with a skewness of less than 2 and a kurtosis of less than 7. Normal distribution would fall within this range, so maximum likelihood estimation was used for model estimation.

Factorial invariance was examined by replicating a study conducted by Chen, Zhang, Raiford, Zhu, and Weiss (2015). In their study, they tested seven levels of nested models to investigate the degree of invariance. Each level had more constraints than the previous one. The steps they employed are listed below.

Before they began testing measurement invariance, Chen et al. fit a baseline model for each group separately. Groups consisted of: Male students of heterogeneous disability groups and female students of heterogeneous disability groups. The factor structure model derived from the referred sample was used as the baseline model for the male and female samples. It included only the primary subtests. Secondary subtests were not included, because most practitioners do not administer the secondary subtests.

The first step was to assess structural invariance by examining the equality of variance-covariance matrices.

The second step in assessing factorial invariance was to test for configural invariance.

The third level was first-order factor-loading invariance, which is also known as metric or weak factorial invariance. Loadings of subtests on each factor were constrained, so that factor loadings were equal across groups. This made the scales of the latent variables the same for both groups, and the unit of measurement is the same.

In the fourth level, intercept invariance or scalar/strong factorial invariance was examined. Any group difference in subtest means resulted from the true mean differences in latent factors. If the subtests had the same latent factor means, they had the same intercepts across groups.

Residual invariance or strict factorial invariance was tested in the fifth level. It examined whether all group differences on the measured variables were attributable to group differences on the common factors. The residuals were composed of unique variance from subtests and measurement error.

The sixth level was second-order factor loading invariance (second-order metric invariance), which assumed that the first-order latent factors increased equally for the same increase in g.

The seventh level tested invariance of disturbances of the first-order factors.

The scale of the latent factors was identified by fixing a factor loading for each factor to one. Several indices of model fit were used to evaluate and compare the models in this study. Single models were evaluated with comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). When using RMSEA, a value less than 0.05 is considered a good fit, and 0.08 is considered acceptable. With

SRMR, values less than 0.08 are considered a good fit. Chen et al. (2015) did not report the SRMR value that indicates acceptable fit. However, according to Kline (2005), values of the SRMR less than 0.10 are generally considered favorable. A value of 0.95 served as the cutoff point for an acceptable fit of all indices ranging from 0 to 1, with 1 being a perfect fit. Competing nested models were evaluated by the change in the chi-square (Δx^2) value. The Akaike Information Criterion (AIC) and sample size adjusted Bayesian Information Criterion (aBIC) were used to compare competing nested and non-nested models. Lower values indicated superior fit.

To determine evidence of invariance, Δx^2 and ΔCFI were utilized jointly. The criterion for rejecting the null hypothesis of invariance was set as a p value of less than 0.001 for the Δx^2 test and an absolute ΔCFI value greater than 0.01 (Chen, Zhang, Raiford, Zhu, and Weiss, 2015). These criteria for rejecting the null hypothesis of invariance were used, because they were utilized in the study that was replicated in this research. Chen, Zhang, Raiford, Zhu, and Weiss (2015) employed these criteria when investigating factor invariance between genders in the standardization sample on the WISC-V. Four of these five researchers were employed by Pearson, the publishers of the WISC-V, and Pearson supports the results of full factorial invariance between genders. Their methods were replicated with a referred sample.

Evaluating Δx^2 is the traditional method of investigating invariance between groups. However, the Δx^2 test has been found to be highly sensitive to sample size and discrepancies from normality (Chen, Zhang, Raiford, Zhu, and Weiss, 2015). In larger sample sizes, trivial differences between groups may be interpreted as noninvariant across populations. Yet, in follow-up diagnostics, unconstraining parameters do not lead to substantially better data-model fit (Kang, McNeish, & Hancock, 2015). For this reason, research has suggested the use of alternative goodness-of-fit indices (GOFs) that are less sensitive to sample size in invariance testing.

Cheung and Rensvold (2002) recommended ΔCFI, which takes the difference of a CFI for an unconstrained model and a constrained model, much like Δx^2. They found that model complexity and sample size had little effect on ΔCFI. They also suggested use of a common cut-off value for ΔCFI that was empirically derived performed well in detecting measurement invariance. Cheung and Rensvold (2002) recommended an absolute ΔCFI value higher than 0.01 (i.e., |ΔCFI| > 0.01) be used as an indicator of a meaningful change in data-model fit. However, Kang, McNeish, and Hancock (2015) indicated that using ΔCFI to assess measurement or structural invariance is not recommended, because, as demonstrated in their simulation study, values suggestive of invariance change significantly as a function of factor loading magnitude, indicators per factor, and sample size. This makes a single cutoff such as that suggested by Cheung and Rensvold difficult to derive, as values obtained and suggested in methodological studies may not generalize well to other data sets. It appears that ΔCFI is superior to Δx^2 with regard to its sensitivity to sample size, model complexity, and overall fit measures, but the problem of deriving a single cutoff score to determine invariance diminishes interpretability of ΔCFI. Therefore, Chen, Zhang, Raiford, Zhu, and Weiss (2015) evaluated ΔCFI and Δx^2 jointly.

Expectations

It was expected that the factor structure of the WISC-V that was derived from the referred sample's data would indicate that the same factors are being measured in the referred sample that were measured in the standardization sample. It was also expected that the model that was derived from the referred sample would be invariant across genders. These expectations are in line with the WISC-V publisher's claim that the instrument is psychometrically sound. However, based on independent researchers' findings, it was not expected that the model would support interpretation beyond the Full Scale IQ. It was expected that most of the score variance

would come from the *g* factor, so interpretation of performance at the subtest level would not

provide meaningful explanations of performance (strengths or weaknesses).

4 RESULTS

First, I examined normality of each subtest and index score using SPSS 24. Skewness ranged from -0.276 to 0.738, while kurtosis ranged from -0.437 to 0.691. Data with skewness greater than -0.8 and less than 0.8 and with kurtosis greater than -3.0 and less than 3.0 are considered normally distributed. Q-Q plots and histograms also suggested that the data are normally distributed. However, the Shapiro-Wilk and the Kolmogorov-Smirnov Tests of Normality indicated that the data are not normally distributed. SPSS recommends using the Shapiro-Wilk Test of Normality only for sample sizes less than 50 and the Kolmogorov-Smirnov Test for larger samples (Davis, 2013). It has been reported that the Kolmogorov-Smirnov Test has low power and should not be used for testing normality (Ghasemi & Zahediasl, 2012). Therefore, for purposes of this study, I determined that the assumption of normality was met.

Table 2

Normality Data of Subtest and Index Scores

Subtest	Skewness	Kurtosis
Similarities	0.273	-0.136
Vocabulary	0.738	0.691
Digit Span	0.239	0.170
Block Design	0.226	-0.061
Visual Puzzles	0.229	0.026
Matrix Reasoning	0.138	-0.437
Figure Weights	0.123	-0.345
Processing Speed	0.261	-0.289
Coding	0.003	-0.379
Symbol Search	-0.065	-0.250

Index Score	Skewness	Kurtosis
Verbal Comprehension	0.472	0.555
Visual Spatial	0.098	0.019
Fluid Reasoning	0.353	-0.066
Working Memory	0.404	-0.008
Processing Speed	-0.276	-0.098

Analyses assessing the measurement properties of the Wechsler Intelligence Scale for Children- Fifth Edition in a sample of n = 473 African-Americans were conducted. The psychometric analyses were divided into two broad categories: 1) dimensionality assessment and 2) assessment of measurement invariance across genders. Results revealed that the second-order confirmatory factor analysis with four hierarchical factors best represented the 10 scale scores from the WISC-V compared to the two-, three-, and five-factor hierarchical models.

Measurement invariance was evaluated using a sequential series of nested model comparisons, based on the four-factor hierarchical structure that best represented the 10 scale scores. Measurement invariance tests supported that the measurement properties of the WISC-V were invariant across males and females in this African-American sample, which provides empirical support for making meaningful comparisons across genders. Comparisons of statistics such as means and regression coefficients can be made, because the measures are comparable across the two groups. The only significant difference that arose was that females had a higher average level on Factor 4, compared to males.

This project employed a two-step analytic strategy for evaluating the measurement properties of the WISC-V in this sample. All models were fit within the confirmatory factor analysis (CFA) framework using Mplus version 7.4 and employing maximum likelihood estimation. As

done by Chen et al. (2015), latent variables were scaled by fixing the factor means at 0 and factor variances at 1. Models were compared using likelihood ratio tests (LRTs), change in CFI (ΔCFI) and standard model fit criteria such as CFI, TLI, and RMSEA. Widely used cut-off values for determining good fit, which were used in this study were CFI > 0.95, TLI > 0.95 (Hu & Bentler, 1999). RMSEA < 0.05 is typically used, but 0.06 was used as the cut-off value because of its use by Chen et al. (2015).

Step 1: Dimensionality Assessment

In Step 1, I evaluated the dimensionality of the WISC-V by comparing two- through five-factor representations for the structural models that were consistent with the hierarchical models presented in Figure 4 of Canivez et al. (2017). Those models are depicted in Table 3.

Table 3

WISC-V Primary Subtest Alignment for CFA Models

Subtests	Two-Factor Hierarchical Model		Three-Factor Hierarchical Model			Four-factor Hierarchical Model				Five-Factor Hierarchical Model				
	Factors		Factors			Factors				Factors				
	F1	F2	F1	F2	F3	F1	F2	F3	F4	F1	F2	F3	F4	F5
Similarities	*		*			*				*				
Vocabulary	*		*			*				*				
Block Design		*		*			*				*			
Visual Puzzles		*		*			*				*			
Matrix Reasoning		*		*			*					*		
Figure Weights		*		*			*					*		
Digit Span	*				*			*					*	
Picture Span		*		*				*					*	
Coding		*			*				*					*
Symbol Search		*			*				*					*

All models include a higher-order general factor.

In my approach, the subtest factors, also called index scores, acted as lower-order (e.g., first-order) factors in a hierarchical model, while a general factor or Full Scale IQ acted as the higher-order factor (e.g., second-order). This is similar to that presented in Chen et al. (2015).

Table 4 provides model fit and model comparison information for the two- through five-factor hierarchical models for this sample of African-American students. Mplus coding for these analyses is provided in Appendix A. Criteria for adequate model fit were CFI and TLI ≥ 0.90, SRMR ≤ 0.09, and RMSEA ≤ 0.08. Criteria for good model fit were CFI and TLI ≥ 0.95 with SRMR and RMSEA ≤ 0.06. Models were considered superior, if they exhibited adequate to good overall fit and displayed meaningfully better fit than alternative models (ΔCFI > 0.01 and ΔRMSEA > 0.015). Akaike Information Criterion (AIC) was considered. AIC does not have a meaningful scale, so the model with the smallest AIC was preferred, because it is the most likely to replicate (Canivez, Watkins, & Dombrowski, 2015).

Results show that the CFAs for the two- and three-factor hierarchical models did not meet criteria for adequate model fit; TLI and CFI ≤ 0.90, RMSEA > 0.08. The four- and five-factor hierarchical model statistics indicated that the data met criteria for good model fit; TLI/CFI ≥ 0.96, SRMR and RMSEA ≤ 0.05. The CFI, TLI, RMSEA, and SRMR were identical for the four- and five-factor hierarchical models. Further comparisons of the two models investigated the LRTs and relative fit indices (AIC/aBIC). The Likelihood Ratio Test suggested that the five-factor hierarchical model did not provide a significant improvement over the four-factor hierarchical model (p = 0.27). AIC and aBIC were slightly smaller for the four-factor hierarchical model. Therefore, due to parsimony, as well as fit indices, the four-factor hierarchical model was considered superior.

Table 4

Factor Structures for the Hierarchical Representations of the WISC-V

	Two-factor hierarchical model	Three-factor hierarchical model	Four-factor hierarchical model	Five-factor hierarchical model
Chi-Square Test of Model Fit	$\chi^2(34) = 245.12$, $p < .0001$	$\chi^2(33) = 159.90$, $p < .0001$	$\chi^2(31) = 66.34$, $p < .0002$	$\chi^2(30) = 67.57$, $p < .0002$
CFI	0.83	0.90	0.97	0.97
TLI	0.78	0.86	0.96	0.96
RMSEA	.12 (.10 - .13)	.09 (.08 - .10)	.05 (.03 - .07)	.05 (.04 - .07)
SRMR	0.07	0.07	0.03	0.03
AIC	21605.34	21522.13	21432.57	21435.80
aBIC	21635.88	21553.66	21466.07	21470.28
Likelihood Ratio Test (Compared to)	---	$\chi^2(1) = 85.22$, $p < .0001$ (Two-Factor)	$\chi^2(2) = 93.46$, $p < .0001$ (Three-Factor)	$\chi^2(1) = 1.23$, $p = .27$ (Four-factor)
Preferred Model	---	Three-Factor	Four-factor	Four-factor
Notes		Factor 2 variance fixed to 0 for convergence		

These fit statistics are very similar to those reported in Table 3 by Canivez, Watkins, and Dombrowski (2017), when they examined model fit with data from the standardization sample. An adapted version of that table with the bifactor model omitted is reported in Table 5.

Table 5

CFA Fit Statistics for WISC-V 10 Primary Subtests

Adapted from Canivez, Watkins, and Dombrowski (2017)

Model[a]	χ^2	*df*	CFI	TLI	SRMR	RMSEA	RMSEA 90% CI	AIC
1 (g)	1296.5	35	.848	.804	.071	.128	[.122, .134]	1226.5
2 (V, P)[b]	1125.9	33	.868	.820	.072	.123	[.117, .129]	1059.9
3 (V, P, and PS)	871.1	32	.899	.858	.062	.109	[.103, .115]	807.1
4 (VC, PR, WM, and PS)	184.9	31	.981	.973	.027	.048	[.041, .054]	122.9
5 (VC, VS, FR, WM, and PS)	134.0	30	.987	.981	.025	.040	[.033, .047]	74.0

Note.   CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean squared residual; RMSEA = root mean square error of approximation; AIC = Akaike's Information Criterion; g = General Intelligence; V = Verbal; P = Performance; PS = Processing Speed; VC = Verbal Comprehension; PR = Perceptual Reasoning; WM = Working Memory; VS = Visual Spatial; FR = Fluid Reasoning;

[a]Model numbers correspond to those reported in the WISC-V Technical and Interpretive Manual and are higher-order models (unless otherwise specified) when more than one first-order factor was specified. [b]Factor 1 (Verbal) and the higher-order factor (g) were linearly dependent on other parameters, so variance estimate was set to zero for model estimation and loss of 1 *df*.

Table 6

Standardized Model Results from Best Fitting Four-factor Model

	Estimate	Standard Error	Estimate/ Standard Error	P-Value
Factor 1 BY				
Similarities	0.85	0.03	26.21	0.000
Vocabulary	0.73	0.03	21.69	0.000
Factor 2 BY				
Block Design	0.31	0.05	6.43	0.000
Visual Puzzles	0.73	0.03	23.48	0.000
Matrix Reasoning	0.67	0.03	20.26	0.000
Figure Weights	0.60	0.04	16.51	0.000
Factor 3 BY				
Digit Span	0.70	0.04	17.76	0.000
Processing Speed	0.61	0.04	15.15	0.000
Factor 4 BY				
Coding	0.70	0.05	15.49	0.000
Symbol Search	0.80	0.05	16.89	0.000
General Factor BY				
Factor 1	0.74	0.04	18.63	0.000
Factor 2	0.91	0.04	24.18	0.000
Factor 3	0.89	0.05	18.40	0.000

Factor 4	0.56	0.05	10.98	0.000
Intercepts				
Similarities	2.21	0.09	25.89	0.000
Vocabulary	2.35	0.09	26.34	0.000
Digit Span	2.45	0.09	26.64	0.000
Block Design	2.57	0.10	26.94	0.000
Visual Puzzles	2.72	0.10	27.29	0.000
Matrix Reasoning	2.47	0.09	26.70	0.000
Figure Weights	3.17	0.11	28.09	0.000
Processing Speed	2.55	0.10	26.90	0.000
Coding	2.16	0.08	25.74	0.000
Symbol Search	2.54	0.09	26.87	0.000
Variances				
General Factor	1.00	0.00	999.00	999.000
Residual Variances				
Similarities	0.28	0.06	5.04	0.000
Vocabulary	0.47	0.05	9.72	0.000
Digit Span	0.51	0.06	9.23	0.000
Block Design	0.91	0.03	31.15	0.000
Visual Puzzles	0.47	0.05	10.20	0.000
Matrix Reasoning	0.55	0.05	12.23	0.000
Figure Weights	0.64	0.04	14.57	0.000

Processing Speed	0.63	0.05	13.07	0.000
Coding	0.51	0.06	7.93	0.000
Symbol Search	0.36	0.08	4.78	0.000
Factor 1	0.45	0.06	7.74	0.000
Factor 2	0.17	0.07	2.41	0.016
Factor 3	0.20	0.09	2.35	0.019
Factor 4	0.69	0.06	12.27	0.000
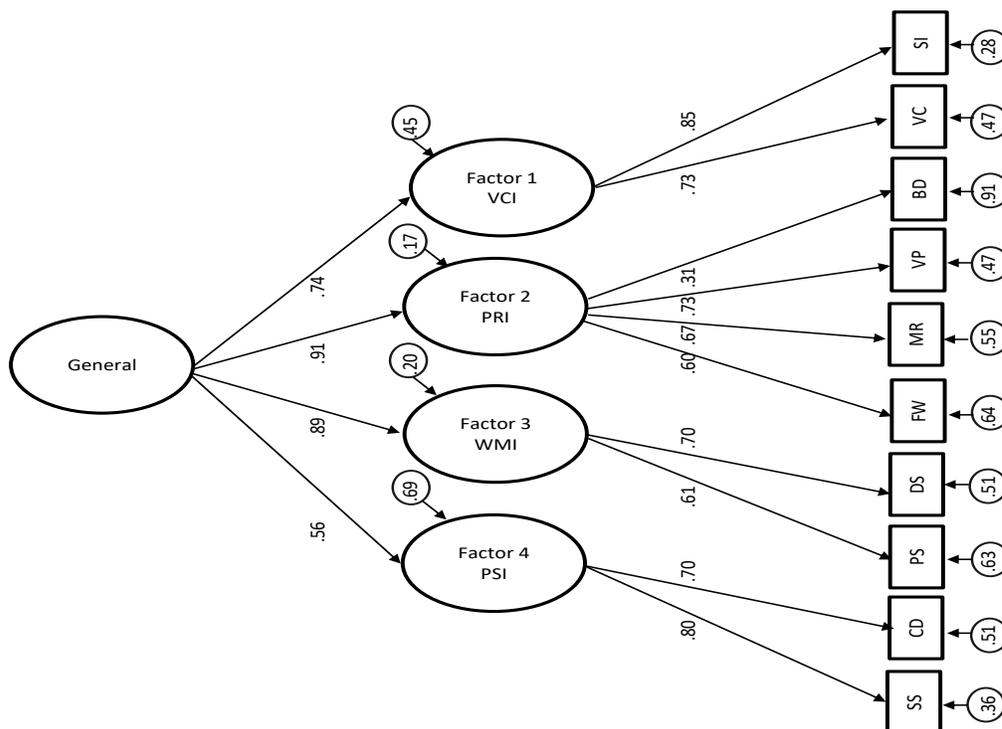
Figure 2   Visual Depiction of Four-factor Hierarchical Model (standardized results)

Table 6 provides standardized model results, while Figure 2 provides a visual depiction

of the four-factor hierarchical model. On the four-factor hierarchical model, the Figure Weights

and Matrix Reasoning subtests were allowed to load onto the same factor as the Visual Puzzles

and Block Design subtests. All of these subtests are thought to be influenced by a latent factor

that can be called Perceptual Reasoning. Figure 2 shows that, with the exception of Block De-

sign, performance on the subtests was heavily influenced by the latent factors described by the

corresponding index. Block Design also loaded poorly on the five-factor hierarchical models

with this sample.

Table 7

Standardized Model Results from Best Fitting Five Factor Model

	Estimate	Standard Error	Estimate/ Standard Error	P-Value
Factor 1 BY				
Similarities	0.86	0.03	26.56	0.000
Vocabulary	0.72	0.03	21.55	0.000
Factor 2 BY				
Block Design	0.32	0.05	6.68	0.000
Visual Puzzles	0.83	0.07	12.54	0.000
Factor 3 BY				
Matrix Reasoning	0.67	0.03	20.49	0.000
Figure Weights	0.59	0.04	16.30	0.000
Factor 4 BY				
Digit Span	0.70	0.04	17.69	0.000
Processing Speed	0.60	0.04	15.01	0.000

Factor 5 BY				
Coding	0.70	0.05	15.18	0.000
Symbol Search	0.80	0.05	16.61	0.000
General Factor BY				
Factor 1	0.72	0.04	18.80	0.000
Factor 2	0.85	0.07	12.15	0.000
Factor 3	1.00	0.00	1068.14	0.000
Factor 4	0.86	0.05	19.10	0.000
Factor 5	0.53	0.05	10.79	0.000
Intercepts				
Similarities	2.21	0.09	25.89	0.000
Vocabulary	2.35	0.09	26.35	0.000
Digit Span	2.45	0.09	26.64	0.000
Block Design	2.57	0.10	26.93	0.000
Visual Puzzles	2.72	0.10	27.29	0.000
Matrix Reasoning	2.47	0.09	26.70	0.000
Figure Weights	3.17	0.11	28.09	0.000
Processing Speed	2.55	0.10	26.90	0.000
Coding	2.16	0.08	25.74	0.000
Symbol Search	2.54	0.09	26.87	0.000
Variances				
General Factor	1.00	0.00	999.00	999.000

Residual Variances				
Similarities	0.27	0.06	4.83	0.000
Vocabulary	0.48	0.05	9.97	0.000
Digit Span	0.51	0.06	9.07	0.000
Block Design	0.90	0.03	28.78	0.000
Visual Puzzles	0.30	0.11	2.74	0.006
Matrix Reasoning	0.55	0.04	12.64	0.000
Figure Weights	0.65	0.04	15.09	0.000
Processing Speed	0.64	0.05	13.09	0.000
Coding	0.51	0.06	7.94	0.000
Symbol Search	0.36	0.08	4.56	0.000
Factor 1	0.48	0.06	8.69	0.000
Factor 2	0.29	0.12	2.43	0.015
Factor 3	0.00	0.00	0.01	0.995
Factor 4	0.27	0.08	3.52	0.000
Factor 5	0.72	0.05	13.53	0.000
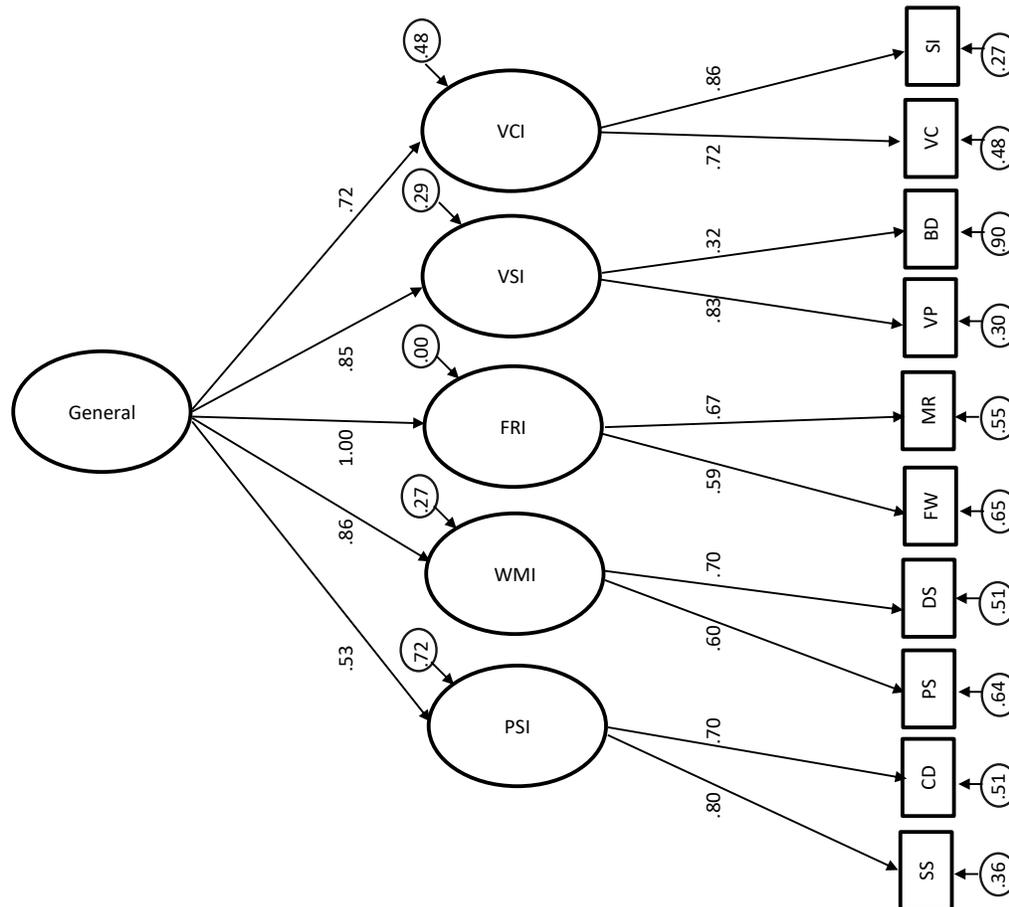
Figure 3   Visual Depiction of Five-Factor Hierarchical Model (standardized results)

Table 7 offers standardized model results, while Figure 3 offers a visual depiction of the

five-factor hierarchical model. On the five-factor hierarchical model, Fluid Reasoning was per-

fectly correlated with the Full Scale IQ. It had a standardized loading of 1.00 on the second-order

g factor. Literature suggests that fluid reasoning factors often show *g* loadings approaching or

reaching 1.00 (Keith et al., 2006; Chen et al., 2015). Again, fluid reasoning has demonstrated to

be the flagship of human cognition. These findings are consistent with those derived from the

standardization sample and published in the *WISC-V Technical and Interpretive Manual*. Fluid

Reasoning had a standardized loading of 0.99 on the second-order g factor with the standardization sample.

This indicates that Fluid Reasoning is measuring the same construct as the Full Scale IQ with this African-American sample. A possible explanation for this phenomenon is that this is a spurious correlation, as Fluid Reasoning is included in the factors that comprise the Full Scale IQ.

Step 2: Evaluating Measurement Invariance

In Step 2, measurement invariance was evaluated based on the factor structure from the best fitting second-order CFA model from Step 1, which was the four-factor model. Measurement invariance was tested using a sequential series of nested model comparisons. I employed a strategy similar to that of Chen et al. (2015), which focused on evaluating invariance at the subtest level, followed by the first-order factor level. Cheung and Rensvold (2002) noted that Likelihood Ratio Tests can often be too sensitive for detecting measurement differences across groups, so I examined changes in CFI across models when making decisions regarding model comparisons. A change of -0.01 in CFI, a frequently used threshold that was utilized by Chen et al. (2015), was considered meaningful.

Table 8 provides model fit details on all steps of the measurement invariance analyses. In phase 1, I verified that the baseline model fit well for both males and females. In phase 2, I formally evaluated measurement invariance in sequential steps. Mplus coding for the analyses is included in Appendix B.

Table 8

Summary of Four-factor Measurement Invariance Analyses

Model	χ² Model Fit	CFI	TLI	RMSEA 95% CI	SRMR	AIC	aBIC	Model Comp	ΔCFI	LRT	Preferred Model
Phase 1: Baseline model fit for each group											
Male (n = 295)	$\chi^2(31)=$ 48.46, p=.02	0.977	0.97	0.04 (.02-.07)	0.04	13446.77	13464.30	---	---	---	---
Female (n = 178)	$\chi^2(31) =$ 53.23, p=.008	0.956	0.94	0.06 (.03-.09)	0.04	7988.76	7989.26	---	---	---	---
Phase 2: Measurement Invariance											
Model 1: Equality of variance-covariance matrices	$\chi^2(65) =$ 100.39, p=.003	0.972	0.96	0.05 (.03-.07)	0.06	21428.23	21492.27				
Model 2: Configural	$\chi^2(62) =$ 101.69, p=.001	0.968	0.95	0.05 (03-.07)	0.04	21435.52	21502.52	---	---	---	---
Model 3: First-order loadings	$\chi^2(68) =$ 106.82, p=.002	0.969	0.96	0.05 (.03-.07)	0.04	21428.65	21489.74	3 vs. 2	0.001	$\chi^2(6) =$ 5.12, p=.53	3
Model 4: First-order load-	$\chi^2(74) =$ 119.92,	0.963	0.96	0.05 (.03-.07)	0.05	21429.76	21484.93	4 vs. 3	-0.006	$\chi^2(6) =$ 13.11,	4*

ings, subtest intercepts	p=.0006									p=.04	
Model 5: First-order loadings, subtest inter-cepts, subtest residual vari-ances	$\chi^2(84) =$ 136.92, p=.0002	0.958	0.96	0.05 (.04-.07)	0.05	21426.75	21472.07	5 vs. 4	-0.005	$\chi^2(10) =$ 17.00, p=.07	5
Model 6: First-order loadings, subtest inter-cepts, subtest residual vari-ances, second-order factor loadings	$\chi^2(87) =$ 139.78, p=.0003	0.958	0.96	0.05 (.03-.07)	0.06	21423.61	21465.98	6 vs. 5	0.000	$\chi^2(3) =$ 2.86, p=.41	6
Model 7: First-order loadings, subtest inter-cepts, subtest residual vari-ances, second-order factor loadings, first-order factor dis-turbances	$\chi^2(91) =$ 144.50, p=.0003	0.957	0.96	0.05 (.03-.07)	0.06	21420.33	21458.76	7 vs. 6	-0.001	$\chi^2(4) =$ 4.72, p=.32	7

Notes: *ΔCFI was small; LRT barely significant

In Model 1, I assessed structural invariance. The variance-covariance matrices were con-

strained to be equal across groups. This set the correlations between variables to be identical in

both groups, which made the means and variances identical for the two groups as well. This

model showed good model fit, providing support for invariant covariance patterns across gen-

ders. Testing for structural invariance is done when the variability of target constructs and/or cor-

relational relationships among them are considered relevant to the generalizability aspect of va-

lidity.

The configural model (Model 2) showed good fit to the data. This provided further sup-

port that the factor structure (e.g., pattern of loadings) was similar for males and females. They

shared the same hierarchical four-factor patterns with corresponding subtests loading on the

same factors.

In Model 3, first-order metric (weak factorial) invariance was tested by constraining the

first-order factor loadings to be equal across groups. With equalized factor loadings, the scales of

the latent variables were the same for males and females, and the unit of measurement was iden-

tical. The likelihood ratio test (p = 0.53) and ΔCFI (0.001) suggested that the model fit did not

significantly worsen with these constraints. This indicated that the magnitude of the relations be-

tween the observed subtest scores and the first-order latent factors were similar for males and

females. Group comparisons of factor variances and covariances are defensible.

In Model 4, I assessed first-order scalar (strong factorial) invariance by constraining the

subtest intercepts, as well as first-order factor loadings, to be equal across genders. According to

the LRT, these constraints significantly reduced model fit (p = 0.04), but the ΔCFI was not sub-

stantial (-0.006). Given the overly sensitive nature of LRTs and the fact that the LRT p-value

barely exceeded the standard alpha of 0.05, I decided that subtest intercepts were equal across males and females and that scalar invariance was met. Chen et al. (2015) also did this when their investigation of scalar invariance suggested reduced model fit, as indicated by the $\Delta\chi^2$, but not the ΔCFI.

In Model 5, I evaluated strict first-order factorial invariance (equal loadings, intercepts, and residual variances) by equating the residual variances of the WISC-V subtests across genders. Both the LRT (p = 0.07) and ΔCFI (-0.005) suggested the more restrictive model with equal residual variances was preferred. This indicated that observed variables were measured with the same precision across groups.

Next, I evaluated group differences in the second-order measurement properties. Model 6 tested the equality of second-order factor loadings across genders (second-order metric invariance). The LRT (p=.41) and ΔCFI (0.000) showed that Model 6, the more parsimonious model with equal second-order loadings across groups, was preferred.

Finally, I tested the equality of the first-order factor disturbances across males and females in Model 7. The LRT (p=.32) and ΔCFI (-0.001) both indicated no significant change in model fit in the more constrained Model 7 when compared to Model 6. Based on Model 7, the means of the first-order factors were fixed at 0 for males and freely estimated for females to be -0.15 for Factor 1, -0.23 for Factor 2, -0.33 for Factor 3, and 0.51 for Factor 4. From a statistical significance standpoint, males and females did not differ for the means of Factors 1, 2, or 3 (p > 0.05 for all), but females did have a significantly greater mean from Factor 4 compared to males (p < 0.001).

Measurement invariance analyses of the four-factor hierarchical model suggested that the measurement properties of the WISC-V operate similarly for male and female African-

Americans, with the only difference occurring on Factor 4, on which females had a significantly higher mean than males. These results provide evidence that WISC-V index and subtest scores of the four-factor hierarchical model have the same meaning for males and females and can be interpreted in the same way.

Table 9

Summary of Five Factor Measurement Invariance Analyses

Model	χ² Model Fit	CFI	TLI	RMSEA 95% CI	SRMR	AIC	aBIC	Model Comp	ΔCFI	LRT	Preferred Model
Phase 1: Baseline model fit for each group											
Male (n = 295)	χ²(30)= 51.74, p=.008	0.971	0.96	0.05 (.02-.07)	0.04	13452.05	13470.10	---	---	---	---
Female (n = 178)	Model would not converge							---	---	---	---
Phase 2: Measure-ment Invari-ance											
Model 1: Equality of variance-covariance matrices	Model would not converge										

In addition to evaluating measurement invariance of the four-factor model across genders, I decided to evaluate the five-factor model across genders as well. I followed the same steps that I employed with the four-factor model. In Phase 1, I assessed baseline model fit for each group. The data for males met criteria for good model fit; TLI/CFI ≥ 0.96, SRMR and RMSEA ≤ 0.05. On the basic within-group CFA, the women-only CFA did not converge to a proper solution. Constraints must be placed for the model to converge, which implies a different factor structure for males versus females (an automatic fail for measurement invariance).

Although there was evidence of measurement variance, I attempted to complete the first step in examining measurement invariance, examining equality of variance-covariance matrices. The model would not converge. Mplus coding for these three analyses is included in Appendix C.

A closer examination of the four- and five-factor models for males revealed that both models provided good fit to the data. Fit statistics are displayed in Table 10. Although it came close, the five-factor model did not provide significant improvement over the four-factor model ($p < 0.07$). Again, this is consistent with findings of independent researchers (Canivez, Watkins, & Dombrowski, 2017; Canivez, Watkins, & Dombrowski, 2015; Dombrowski et al., 2015) that found a lack of empirical support for the five-factor model of the WISC-V.

Table 10

Comparison of Factor Structures of the WISC-V for Males

	Four-factor hierarchical model	Five-factor hierarchical model
Chi-Square Test of Model Fit	$\chi^2(31) = 48.46$, p = .02	$\chi^2(30) = 51.74$, p = .008
CFI	0.977	0.971
TLI	0.97	0.96
RMSEA	0.04 (.02-.07)	0.05 (.02-.07)
SRMR	0.04	0.04
AIC	13446.77	13452.05
aBIC	13464.30	13470.10
Likelihood Ratio Test (Compared to)		$\chi^2(1) = 3.28$, p = .07 (Four-factor)
Preferred Model		Four-factor
Notes		

Possible reasons for this nonconvergence are: 1) The model is not appropriate for the data. This is feasible, given that in the comparison of factor structures, the four-factor structure is preferred over the five-factor structure. In the five-factor hierarchical model, Factor 3 (which is the product of splitting up the four observed subtests defining the second factor of the four-factor solution) seems to be causing problems with the convergence. Again, this could be consistent with the hypothesis that these four subtests should not be separated into two distinct factors. 2)

Computer simulation studies have revealed that nonconvergence or improper solutions are more likely to occur when CFA models have only two indicators per factor, which is the case with this model. At least three indicators per factor are recommended. 3) These models may be too complicated for the data at hand. The factor structure is complicated, and by exploring multiple groups, the number of estimated parameters increases. It is suggested that minimum sample size is at least 10 times the number of free parameters, but a 20:1 ratio would be better. The baseline model for females estimated 30 parameters, which may interfere with this sample of less than n = 200 subjects meeting the analytical demands (Kline, 2005).

The final question posed in the study, Does the factor structure derived from a referred sample support interpreting the FSIQ, index scores, and subtest scores of the WISC-V or should interpretations be based on the FSIQ only?, can be answered by examining the reliabilities and proportions of variance attributed to the subtest scores, index scores, and FSIQ.

The *WISC-V Technical and Interpretive Manual* (Wechsler, 2014) recommends interpreting results of this assessment by reporting and describing the Full Scale IQ score. Next, each of the factor index scores is reported and described. The third step involves comparisons of the primary (factor) index scores to the Full Scale IQ. A significant difference between the two scores represents a strength or weakness in that cognitive area. Next, factor index level pairwise comparisons are made, followed by evaluation of subtest-level strengths and weaknesses. Finally, subtest-level pairwise comparisons are made. In order to make these comparisons, the subtests, indexes, and Full Scale IQ must measure the constructs that they purport to measure.

Reliability coefficients reveal the degree to which scores are free from random error. Generally, reliability coefficients around 0.90 are considered excellent, coefficients around 0.80 are very good, and values around 0.70 are adequate. According to Kline (2005), that means that

the scores account for 90%, 80%, and 70% of the variance in the constructs they are intended to measure. As reliability coefficients approach zero, the scores are more like random numbers, which measure nothing in particular.

Reliability is the proportion of true variance relative to total variance (true variance plus error variance). Reliability and the proportion of variance of a measured variable are calculated through squared multiple correlation (SMC), where the measured variable (subtest score) is the dependent variable and the factor (index score) is the independent variable. $\text{SMC}_{\text{var i}} = \frac{\lambda_i^2}{\lambda_i^2 + \theta_{ii}}$. SMC is calculated by squaring the factor loading of a variable and dividing that value by itself plus the residual variance associated with the variable. The proportion of variance in the set of variables accounted for by a factor is also equal to the sum of squared factor loadings for the factor divided by the number of variables in that factor (Tabachnick & Fidell, 2013). These values are shown in Tables 11 and 12.

Table 11

Squared Multiple Correlations of the WISC-V Factor Indexes

Factor Indexes	Squared Multiple Correlation
Verbal Comprehension	0.52
Visual Spatial	0.71
Fluid Reasoning	1.00
Working Memory	0.73
Processing Speed	0.28

The Full Scale IQ or *g* factor accounts for 65% of the total variance. Table 8 shows that 52% of the variance in Verbal Comprehension is accounted for by the Full Scale IQ. The Full Scale IQ accounts for 71% of the variance in the Visual Spatial Index, 100% of the variance in Fluid Reasoning, 73% of the variance in Working Memory, and 28% of the variance in Processing Speed. This indicates that the Verbal Comprehension and Processing Speed Indexes are not adequate indicators of Full Scale IQ. Therefore, WISC-V scores should not be interpreted at the index level.

Table 12

Proportions of Variance in the WISC-V 10 Primary Subtests

Subtests	Factor 1		Factor 2		Factor 3		Factor 4		Factor 5	
	b	s^2	b	s^2	b	s^2	b	s^2	b	s^2
Similarities	0.86	0.74								
Vocabulary	0.72	0.52								
Block Design			0.32	0.10						
Visual Puzzles			0.83	0.69						
Matrix Reasoning					0.67	0.45				
Figure Weights					0.59	0.35				
Digit Span							0.70	0.49		
Picture Span							0.60	0.36		
Coding									0.70	0.49
Symbol Search									0.80	0.64
Sum of Squared Loadings										
	1.26		0.79		0.80		0.85		1.13	
Proportion of Variance	0.63		0.40		0.40		0.43		0.57	

| Proportion of Covariance | 0.26 | 0.16 | 0.17 | 0.18 | 0.23 |

b = loading of subtest on factor S^2 = variance in subtest explained by the factor

Reliability coefficients of the subtests ranged from 0.10 to 0.74. The Similarities subtest was the only one with a reliability coefficient (0.74) within the range that suggests that it adequately measures the constructs of the factor on which it loads. All other reliability coefficients were too low for interpretation. Some even suggest, as they approach zero, that the subtest is not measuring any particular construct. Therefore, WISC-V results should not be interpreted at the subtest level.

Dombrowski et al. (2015) stated that low reliability coefficients at the subtest level suggest that little interpretive weight should be placed on index scores derived from these subtests, because little true score variance exists at the group level that is independent of the general factor. Again, reiterating the belief that the WISC-V should not be interpreted beyond the general factor or Full Scale IQ.

The first factor, Verbal Comprehension, accounts for 26 percent of the variance in FSIQ. Factor 2, Visual Spatial, accounts for 16 percent, and Factor 3, Fluid Reasoning, accounts for 17 percent of the variance in the Full Scale IQ. Factor 4, Working Memory, explains 18 percent, and Factor 5, Processing Speed, explains 23 percent of the variance in the Full Scale IQ.

Sixty-three percent of the variance in the set of variables, Similarities and Vocabulary, is accounted for by Verbal Comprehension. Visual Spatial and Fluid Reasoning each account for 40 percent of the variance in the set of variables that comprise those factors. Factor 4, Working Memory, accounts for 43 percent of the variance in Digit Span and Picture Span, while Factor 5, Processing Speed, accounts for 57 percent of the variance in Coding and Symbol Search.

These results are consistent with previous findings (Dombrowski, Canivez, Watkins, & Beaujean, 2015; Canivez, Watkins, & Dombrowski, 2015, 2017). Most of the WISC-V variance was contributed by *g* and the reliability coefficients of the subtests were low. According to the *Standards* (AERA et al., 2014), interpretation of subscores requires demonstration of the scores' "distinctiveness and reliability" (Standard 1.14). These requirements are not met with the WISC-V, suggesting that interpretation of the WISC-V should be done at the Full Scale IQ level only.

5 DISCUSSION

Conclusions

The purpose of this study was to answer the following questions: Does the WISC-V measure the same constructs for a sample of African-American students who have been referred for evaluation as compared to the standardization sample? Will a confirmatory factor analysis of data gathered from 10 subtests of a referred sample fit the factor structure published in the WISC-V manual? Is the factor structure of the WISC-V invariant across genders in a referred sample? Does the factor structure derived from a referred sample support interpreting the FSIQ, index scores, and subtest scores of the WISC-V or should interpretations be based on the FSIQ only?

The results showed that the WISC-V measures the same constructs for a sample of African-American students who have been referred for evaluation as it does for the standardization sample. However, there was stronger empirical support for a hierarchical factor representation for the WISC-V with four latent constructs than there was for the five latent constructs. Data from this sample of African-Americans fit both models well, with the fit to the four-factor hierarchical model being deemed slightly superior. Its superiority was determined, because goodness-of-fit indices (LRT ($p = 0.27$) and ΔCFI $= 0.00$) did not support the use of a more complex model over the more parsimonious model.

The Block Design subtest loaded poorly on the four-factor hierarchical model with a factor loading of 0.31 and on the five-factor hierarchical model with a factor loading of 0.32. This is inconsistent with this subtest's factor loading (0.76) on the five-factor hierarchical model published in the manual.

On the five-factor hierarchical model, Fluid Reasoning is measuring the same construct as the Full Scale IQ with this African-American sample. Its perfect correlation suggests that it is empirically redundant and can be eliminated from the structural model. These results are consistent with those presented by the test publishers as well as independent researchers.

Measurement invariance tests revealed that the four-factor hierarchical model was invariant across genders, with females scoring slightly higher on Factor 4 than males. Data from the males in the referred sample fit the five-factor hierarchical model published in the WISC-V manual; however, data from the females did not fit the model. This indicated that the five-factor hierarchical model was not invariant across genders or that there may be a problem related to the data, such as sample size, or the design of the structural model, which has only two indicators per factor.

When evaluating the interpretability of WISC-V results, reliability coefficients and proportions of variance attributed to the subtest scores, index scores, and Full Scale IQ were examined. The vast majority of the total variance (65%) was accounted for by the Full Scale IQ. Verbal Comprehension and Processing Speed were not adequate indicators of Full Scale IQ, so interpretations at the index level are not supported. With the exception of the Similarities subtest, reliability coefficients of subtest scores were too low for interpretation. Therefore, interpretation should remain at the Full Scale IQ level. No valid inferences about strengths and weaknesses can be made. The model does not support interpretation methods that are recommended in the manual. Additionally, the five-factor model is the one employed by the test publishers and the one from which inferences about the examinees are being made. However, that model is not supported by data from the females in this sample. Therefore, inferences about African-American fe-

males within this sample, even at the Full Scale IQ level, are not supported by the current five-factor structural model.

The question of why replication studies are not being conducted has become a source of crisis-level anxiety among psychologists. Some reasons given for avoiding replication studies are lack of prestige associated with replications and difficulty conclusively stating whether the replication confirms or disconfirms previous findings (Earp & Trafimow (2015). Earp and Trafimow (2015) stated that replications do not need to be conclusive to be informative. To increase the informativeness of replication attempts, the researcher should utilize the following techniques: 1) carefully define the effects and methods that he/she intend to replicate; 2) follow as exact as possible the methods of the original study; 3) have an adequate sample size to detect an effect, if one is present; 4) make complete details about the replication available, so others may evaluate the replication attempt or attempt another replication; and 5) evaluate the replication results. All of these techniques were followed in this study. Therefore, results of this replication study may increase confidence in the validity of the findings by researchers who embraced the four-factor hierarchical model as the best model to demonstrate the concepts being measured by the Wechsler scales (Watkins, Wilson, Kotz, Carbone, & Babula, 2006; Bodin, Pardini, Burns, and Stevens, 2009; Chen & Zhu, 2012; Nakano & Watkins, 2013; Chen, Zhang, Raiford, Zhu, and Weiss, 2015; Canivez, Watkins, & Dombrowski, 2015).

Implications

During the 2014-2015 school year, 6.6 million students enrolled in public schools in the United States received special education and related services. School psychologists serve as integral members of the multidisciplinary teams that determine eligibility for special education and related services, because we administer the psycho-educational evaluations used in making deci-

sions about eligibility. Benson et al. (2019) examined test usage and assessment procedures of school psychologists. The Wechsler Intelligence Scale for Children- Fifth Edition was determined to be the most frequently administered cognitive assessment. It was used more frequently than the next five most used cognitive tests combined. Eighty percent of the 1317 school psychologists participating in the study utilize the WISC-V. On average, it was administered 3.49 times per month by each of the school psychologists. If these results are generalized to the approximately 32,300 school psychologists practicing in this country's public schools, it would be expected that the WISC-V would be administered 90,182 times per month. That provides 90,182 opportunities for students' lives to be changed, based on the interpretations of the WISC-V results. It is imperative that this instrument is interpreted appropriately.

Suggestions for Further Research

These findings provide some evidence of structural validity of the WISC-V with a referred sample of African-American students. However, more research is needed to gain further support of the WISC-V's structural validity with this subpopulation of students.

It may be beneficial to replicate this study with a national sample of African-American students. This would reveal whether the results of this study are unique to this sample of students, who attended the same school system, or whether they are generalizable to the African-American population.

It is suggested that empirical investigations of the predictive validity of the WISC-V with African-American students are conducted. The WISC-V manual discussed the relations of the WISC-V with several external instruments by examining scores of nonclinical and clinical samples. None of those samples were strictly African-American. An investigation into the predictive

validity of the WISC-V with African-American students may provide more support for its validity as an instrument that should be used with this subpopulation.

Future researchers may want to investigate the usefulness of the Block Design subtest with this population. Block Design is the only subtest that loads onto the Visual Spatial Index that is used in the calculation of the Full Scale IQ. However, with this sample, its correlation with the Visual Spatial factor was 0.32 on the five-factor hierarchical model. Visual Puzzles may be the better choice for inclusion in the Full Scale IQ, because its correlation with the Visual Spatial factor was 0.83 for this sample of African-Americans.

It is suggested that further investigation into the reason that African-American females scored significantly higher on processing speed tasks than males is conducted. This research could help determine whether this significant difference in performance is actually meaningful.

Studies that employ exploratory factor analysis and/or bifactor models to examine the structural validity of the WISC-V with African-American students are strongly encouraged. By eliminating the preconceived beliefs inherit in confirmatory factor analysis, exploratory factor analysis will present the researchers with the freedom to assess more combinations of constructs that the WISC-V is possibly measuring, while bifactor models will allow them to find the sources of variability in the models, which may aid in interpretability of the WISC-V with this subpopulation.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Center on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Ary, D., Jacobs, L. C., & Razavieh, A. (1996). *Introduction to research in education.* (5th ed.). Fort Worth, Texas: Harcourt BraceCollege Publishers.

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology, 72,* 29-48. doi: 10.1016/j.jsp.2018.12.004

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*(5), 881-906. doi: 10.1037/a0023222

Betts, J., Reschley, A., Pickart, M., Heistad, D., Sheran, C., & Marston, D. (2008). An examination of predictive bias for second grade reading outcomes from measures of early literacy skills in kindergarten with respect to English-language learners and ethnic subgroups. *School Psychology Quarterly, 23*(4), 553-570.

Bodin, D., Pardini, D. A., Burns, T. G., & Stevens, A. B. (2009). Higher order factor structure of the WISC-IV in a clinical neuropsychological sample. *Child Neuropsychology, 15,* 417-424.

Canivez, G. L. (2014). Structural validity of the WISC-IV with a referred sample: Direct versus indirect hierarchical structures. *School Psychology Quarterly, 29*, 38-51. http://dx.doi.org/10.1037.spq0000032

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2015, November 16). Factor structure of the Wechsler Intelligence Scale for Children- Fifth Edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0000238

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children- Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 29*(4), 458-472. http://dx.doi.org/10.1037/pas0000358

Chen, H., Zhang, O., Raiford, S. E., Zhu, J., & Weiss, L. G. (2015). Factor invariance between genders on the Wechsler Intelligence Scale for Children–Fifth Edition. *Personality and Individual Differences, 86*, 1-5. doi:10.1016/j.paid.2015.05.020

Chen, H., & Zhu, J. (2012). Measurement invariance of WISC-IV across normative and clinical samples. *Personality and Individual Differences, 52*, 161-166. doi:10.1016/j.paid.2011.10.006

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of Validity Evidence for Educational and Psychological Tests: A Follow-Up Study. *Educational and Psychological Measurement,70*(5), 732-743. doi: 10.1177/0013164410379323

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68, 397-412. doi:10.1177/0013164407310130

Cokley, K. & Moore, P. (2007). Moderating and mediating effects of gender and psychological disengagement on the academic achievement of African-American college students. *Journal of Black Psychology, 33,* 169-187.

Davis, C. (2013). *SPSS for Applied Sciences: Basic Statistical Testing*. Collingwood, Vic: CSIRO PUBLISHING. Retrieved from http://ezproxy.gsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=633778&site=eds-live&scope=site

Devena, S. E., Gay, C. E., & Watkins, M. W. (2013). Confirmatory factor analysis of the WISC-IV in a hospital referral sample. *Journal of Psychoeducational Assessment, 31*(6), 591-599. doi: 10.1177/0734282913483981

Dimitrov, D. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149. doi: 10.1177/0748175610373459

Dombrowski, S. C. (2014). Investigating the structure of the WJ-III cognitive in early school age through two exploratory bifactor analysis procedures. *Journal of Psychoeducational Assessment, 32*, 483-494.

Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Alexander Beaujean, A. (2015). Exploratory bifactor analysis of the Wechsler Intelligence Scale for Children—Fifth Edition with the 16 primary and secondary subtests. *Intelligence*, 53, 194-201. doi: 10.1016/j.intell.2015.10.009

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*, 621-631.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism, 10*(2), 486-489. doi: 10.5812/ijem.3505

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117-135. doi: http://dx.doi.org/10.1016/j.jsp.2006.05.005

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219-230.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 6,* 537-549. doi: 10.1007/s11336-001-9218-4

Kang, Y., McNeish, D. M., & Hancock, G. R. (2015). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement*, 1-29. doi: 10.1177/0013164415603764

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children- Fourth Edition: What does it measure? *School Psychology Review, 35,* 108-127.

Kline, W. C. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Mandara, J., Varner, F., & Richman, S. (2010). Do African-American mothers really "love" their sons and "raise" their daughters? *Journal of Family Psychology, 24*(1), 41-50.

McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment, 26*(1), 207-214. doi:10.1037/a0034832

McMillian, M. M., Carr, M., Hodnett, G., & Campbell, F. A. (2016). A longitudinal study of academic identification among African-American males and females. *Journal of Black Psychology, 42*(6), 508-529. doi: 10.1177/0095798415603845

McMillian, M. M., Frierson, H. T. & Campbell, F. A. (2011). Do gender differences exist in the academic identification of African-American elementary school-aged children? *Journal of Black Psychology, 37*(1), 78-98. doi: 10.1177/0095798410366709

Nakano, S. & Watkins, M. W. (2013). Factor structure of the Wechsler Intelligence Scales for Children- Fourth Edition among referred native American students. *Psychology in the Schools, 50*(10), 957-968.

Ogbu, J. U. (2003). *Black American students in an affluent suburb: A study of academic disengagement*. Mahwah, NJ: Lawrence Erlbaum.

Osborne, J. W. (1995). Academics, self-esteem, and race: A look at the underlying assumptions of the disidentification hypothesis. *Personality and Social Psychology Bulletin, 21,* 449-455.

Oren, C., Kennet-Cohen, T., Turvall, E., & Allalouf, A. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema, 26*(1), 117-126. Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*(5), 881-906. doi:10.1037/a0023222

Padilla, J. & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136-144. doi: 10.7334/psicothema2013.259

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696.

Rios, J. & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108-116. doi: 10.7334/psicothema2013.260

Sireci, S. & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100-107. doi: 10.7334/psicothema2013.256

Skiba, R. (2013). CCBD'S position summary on federal policy on disproportionality in special education. *Behavior Disorders, 38*(2), 108-120.

Skinner, O. D., Perkins, K., Wood, D. & Kurtz-Costes, B. (2016). Gender development in African-American youth. *Journal of Black Psychology, 42*(5), 394-423.

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics - sixth edition.* Boston, MA: Pearson Education, Inc.

Varner, F. & Mandara, J. (2013). Differential parenting of African-American adolescents as an explanation for gender disparities in achievement. *Journal of Research on Adolescence, 24*(4), 667-680.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children- Fourth Edition among a national sample of referred students. *Psychological Assessment, 22,* 782-787.

Watkins, M. W. (2013). Omega. *[Computer software].* Phoenix, AZ: Ed & Psych Associates.

Watkins, M. W., Wilson, S. M., Kotz, K. M., Carbone, M. C. & Babula, T. (2006). Factor structure of the Wechsler Intelligence Scale for Children- Fourth Edition among referred students. *Educational and Psychological Measurement, 66,* 975-983.

Wechsler, D. (2003). Wechsler Intelligence Scale for Children- Fourth Edition, Technical and

    interpretive manual. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2014). Wechsler Intelligence Scale for Children- Fifth Edition, Technical and

    interpretive manual. Bloomington, MN: Pearson.

Zhang, D., Katsiyannis, A., Ju, S., & Roberts, E. (2014). Minority Representation in Special

    Education: 5-Year Trends. *Journal Of Child & Family Studies*, *23*(1), 118-127.

    doi:10.1007/s10826-012-9698-6

APPENDICES

Appendix A

Dimensionality Assessment

TITLE: DC- Rachel - Two-factor Hierarchical Model

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

Factor1 by si_ss* vc_ss ds_ss;

Factor1@1;

Factor2 by bd_ss* vp_ss mr_ss fw_ss ps_ss cd_ss ss_ss;

Factor2@1;

GenFac by Factor1* (1);

GenFac by Factor2* (1);

GenFac@1;

OUTPUT:

 STDYX;

TITLE: DC- Rachel – Three-factor Hierarchical Model

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

Factor1 by si_ss vc_ss;

Factor2 by bd_ss vp_ss mr_ss fw_ss ps_ss;

Factor2@0;

Factor3 by ds_ss cd_ss ss_ss;

GenFac by Factor1 Factor2 Factor3;

OUTPUT:

 STDYX;

TITLE: Rachel Dissertation – Four-factor Hierarchical Model

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

Factor1 by si_ss* vc_ss;

Factor1@1;

Factor2 by bd_ss* vp_ss mr_ss fw_ss;

Factor2@1;

Factor3 by ds_ss* ps_ss;

Factor3@1;

Factor4 by cd_ss* ss_ss;

Factor4@1;

GenFac by Factor1* Factor2 Factor3 Factor4;

GenFac@1;

OUTPUT:

STDYX;

TITLE: Rachel Dissertation – Five-factor Hierarchical Model

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

Factor1 by si_ss vc_ss;

Factor2 by bd_ss vp_ss;

Factor3 by mr_ss fw_ss;

Factor3 (v);

Factor4 by ds_ss ps_ss;

Factor5 by cd_ss ss_ss;

GenFac by Factor1 Factor2 Factor3 Factor4 Factor5;

OUTPUT:stdyx;

 MODEL CONSTRAINT:

 NEW (v0);

 v = exp(v0);

Appendix B

Measurement Invariance

TITLE: DC- Rachel - Invariance – Baseline Model for Black Females

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 MISSING = ALL (-999);

useobs = male == 0;

ANALYSIS: ESTIMATOR = ML;

MODEL:

Factor1 by si_ss* vc_ss;

Factor1@1;

Factor2 by bd_ss* vp_ss mr_ss fw_ss;

Factor2@1;

Factor3 by ds_ss* ps_ss;

Factor3@1;

Factor4 by cd_ss* ss_ss;

Factor4@1;

GenFac by Factor1* Factor2 Factor3 Factor4;

GenFac@1; OUTPUT: STDYX;

TITLE: DC- Rachel - Invariance – Baseline Model for Males

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 MISSING = ALL (-999);

useobs = male == 1;

ANALYSIS: ESTIMATOR = ML;

MODEL:

Factor1 by si_ss* vc_ss;

Factor1@1;

Factor2 by bd_ss* vp_ss mr_ss fw_ss;

Factor2@1;

Factor3 by ds_ss* ps_ss;

Factor3@1;

Factor4 by cd_ss* ss_ss;

Factor4@1;

GenFac by Factor1* Factor2 Factor3 Factor4;

GenFac@1;

OUTPUT:

STDYX;

TITLE: DC- Rachel - Invariance – Model 1

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 GROUPING = male (0=Women 1=Men);

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

si_ss (1); vc_ss (2); ds_ss (3); bd_ss (4); vp_ss (5); mr_ss (6); fw_ss (7);

 ps_ss (8); cd_ss (9); ss_ss (10);

[si_ss] (91); [vc_ss] (92); [ds_ss] (93); [bd_ss] (94); [vp_ss] (95); [mr_ss] (96);

[fw_ss] (97); [ps_ss] (98); [cd_ss] (99); [ss_ss] (910);

si_ss with vc_ss (20); si_ss with ds_ss (21);si_ss with bd_ss (22);si_ss with vp_ss (23);

si_ss with mr_ss (24);si_ss with fw_ss (25);si_ss with ps_ss (26); si_ss with cd_ss (27);

si_ss with ss_ss (28);

vc_ss with ds_ss (29); vc_ss with bd_ss (30);vc_ss with vp_ss (31);vc_ss with mr_ss

(32);

vc_ss with fw_ss (33);vc_ss with ps_ss (34);vc_ss with cd_ss (35);vc_ss with ss_ss (36);

ds_ss with bd_ss (37);ds_ss with vp_ss (38);ds_ss with mr_ss (39);ds_ss with fw_ss (40);

ds_ss with ps_ss (41);ds_ss with cd_ss (42);ds_ss with ss_ss (43);

bd_ss with vp_ss (44);bd_ss with mr_ss (45);bd_ss with fw_ss (46);

bd_ss with ps_ss (47);bd_ss with cd_ss (48);bd_ss with ss_ss (49);

vp_ss with mr_ss (50);vp_ss with fw_ss (51);

vp_ss with ps_ss (52);vp_ss with cd_ss (53);vp_ss with ss_ss (54);

mr_ss with fw_ss (55);mr_ss with ps_ss (56);mr_ss with cd_ss (57);mr_ss with ss_ss (58);

fw_ss with ps_ss (59);fw_ss with cd_ss (60);fw_ss with ss_ss (61);

ps_ss with cd_ss (62);ps_ss with ss_ss (63);

cd_ss with ss_ss (64);

MODEL WOMEN:

si_ss (1); vc_ss (2); ds_ss (3); bd_ss (4); vp_ss (5); mr_ss (6); fw_ss (7);

 ps_ss (8); cd_ss (9); ss_ss (10);

[si_ss] (91); [vc_ss] (92); [ds_ss] (93); [bd_ss] (94); [vp_ss] (95); [mr_ss] (96);

[fw_ss] (97); [ps_ss] (98); [cd_ss] (99); [ss_ss] (910);

si_ss with vc_ss (20); si_ss with ds_ss (21);si_ss with bd_ss (22);si_ss with vp_ss (23);

si_ss with mr_ss (24);si_ss with fw_ss (25);si_ss with ps_ss (26); si_ss with cd_ss (27);

si_ss with ss_ss (28);

vc_ss with ds_ss (29); vc_ss with bd_ss (30);vc_ss with vp_ss (31);vc_ss with mr_ss

(32);

vc_ss with fw_ss (33);vc_ss with ps_ss (34);vc_ss with cd_ss (35);vc_ss with ss_ss (36);

ds_ss with bd_ss (37);ds_ss with vp_ss (38);ds_ss with mr_ss (39);ds_ss with fw_ss (40);

ds_ss with ps_ss (41);ds_ss with cd_ss (42);ds_ss with ss_ss (43);

bd_ss with vp_ss (44);bd_ss with mr_ss (45);bd_ss with fw_ss (46);

bd_ss with ps_ss (47);bd_ss with cd_ss (48);bd_ss with ss_ss (49);

vp_ss with mr_ss (50);vp_ss with fw_ss (51);

vp_ss with ps_ss (52);vp_ss with cd_ss (53);vp_ss with ss_ss (54);

mr_ss with fw_ss (55);mr_ss with ps_ss (56);mr_ss with cd_ss (57);mr_ss with ss_ss (58);

fw_ss with ps_ss (59);fw_ss with cd_ss (60);fw_ss with ss_ss (61);

ps_ss with cd_ss (62);ps_ss with ss_ss (63);

cd_ss with ss_ss (64);

OUTPUT:

 STDYX;

TITLE: DC- Rachel - Invariance - Model 2

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 GROUPING = male (0=Women 1=Men);

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2@1;[Factor2@0];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3@1;[Factor3@0];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4@1;[Factor4@0];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

MODEL WOMEN:

! Factor loadings all estimated

Factor1 by si_ss* (91);

Factor1 by vc_ss* (92);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (93);

Factor2 by vp_ss* (94);

Factor2 by mr_ss* (95);

Factor2 by fw_ss* (96);

Factor2@1;[Factor2@0];

Factor3 by ds_ss* (97);

Factor3 by ps_ss* (98);

Factor3@1;[Factor3@0];

Factor4 by cd_ss* (99);

Factor4 by ss_ss* (100);

Factor4@1;[Factor4@0];

! second order loadings;

GenFac by Factor1* (101);

GenFac by Factor2 (102);

GenFac by Factor3 (103);

GenFac by Factor4 (104);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

OUTPUT:

STDYX;

TITLE: DC- Rachel - Invariance - Model 3

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 GROUPING = male (0=Women 1=Men);

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2@1;[Factor2@0];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3@1;[Factor3@0];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4@1;[Factor4@0];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

MODEL WOMEN:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1*;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2*;[Factor2@0];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3*;[Factor3@0];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4*;[Factor4@0];

! second order loadings;

GenFac by Factor1* (101);

GenFac by Factor2 (102);

GenFac by Factor3 (103);

GenFac by Factor4 (104);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 OUTPUT:

 STDYX;

TITLE: DC- Rachel - Invariance - Model 4

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

ps_ss cd_ss ss_ss;

GROUPING = male (0=Women 1=Men);

MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2@1;[Factor2@0];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3@1;[Factor3@0];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4@1;[Factor4@0];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

ps_ss cd_ss ss_ss;

MODEL WOMEN:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1*;[Factor1*];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2*;[Factor2*];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3*;[Factor3*];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4*;[Factor4*];

! second order loadings;

GenFac by Factor1* (101);

GenFac by Factor2 (102);

GenFac by Factor3 (103);

GenFac by Factor4 (104);

GenFac@1;[GenFac@0];

! item intercepts;

![si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

! ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

ps_ss cd_ss ss_ss;

OUTPUT:

STDYX;

TITLE: DC- Rachel - Invariance - Model 5

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 GROUPING = male (0=Women 1=Men);

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2@1;[Factor2@0];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3@1;[Factor3@0];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4@1;[Factor4@0];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

  si_ss (51); vc_ss (52); ds_ss (53); bd_ss (54);

  vp_ss (55); mr_ss (56); fw_ss (57);

  ps_ss (58); cd_ss (59); ss_ss (60);

MODEL WOMEN:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1*;[Factor1*];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2*;[Factor2*];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3*;[Factor3*];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4*;[Factor4*];

! second order loadings;

GenFac by Factor1* (101);

GenFac by Factor2 (102);

GenFac by Factor3 (103);

GenFac by Factor4 (104);

GenFac@1;[GenFac@0];

! item intercepts;

![si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

! ps_ss cd_ss ss_ss];

! Item variances

 si_ss (51); vc_ss (52); ds_ss (53); bd_ss (54);

 vp_ss (55); mr_ss (56); fw_ss (57);

 ps_ss (58); cd_ss (59); ss_ss (60);

 OUTPUT:

 STDYX;

TITLE: DC- Rachel - Invariance - Model 6

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

ps_ss cd_ss ss_ss;

GROUPING = male (0=Women 1=Men);

MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2@1;[Factor2@0];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3@1;[Factor3@0];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4@1;[Factor4@0];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

ps_ss cd_ss ss_ss];

! Item variances

si_ss (51); vc_ss (52); ds_ss (53); bd_ss (54);

vp_ss (55); mr_ss (56); fw_ss (57);

ps_ss (58); cd_ss (59); ss_ss (60);

MODEL WOMEN:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1*;[Factor1*];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2*;[Factor2*];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3*;[Factor3*];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4*;[Factor4*];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac*;[GenFac@0];

! item intercepts;

![si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

! ps_ss cd_ss ss_ss];

! Item variances

si_ss (51); vc_ss (52); ds_ss (53); bd_ss (54);

vp_ss (55); mr_ss (56); fw_ss (57);

ps_ss (58); cd_ss (59); ss_ss (60);

OUTPUT:

STDYX;

TITLE: DC- Rachel - Invariance - Model 7

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

 GROUPING = male (0=Women 1=Men);

 MISSING = ALL (-999);

ANALYSIS: ESTIMATOR = ML;

MODEL:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2@1;[Factor2@0];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3@1;[Factor3@0];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4@1;[Factor4@0];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

  si_ss (51); vc_ss (52); ds_ss (53); bd_ss (54);

  vp_ss (55); mr_ss (56); fw_ss (57);

   ps_ss (58); cd_ss (59); ss_ss (60);

MODEL WOMEN:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1*];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2 by mr_ss* (5);

Factor2 by fw_ss* (6);

Factor2@1;[Factor2*];

Factor3 by ds_ss* (7);

Factor3 by ps_ss* (8);

Factor3@1;[Factor3*];

Factor4 by cd_ss* (9);

Factor4 by ss_ss* (10);

Factor4@1;[Factor4*];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac*;[GenFac@0];

! item intercepts;

![si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

! ps_ss cd_ss ss_ss];

! Item variances

 si_ss (51); vc_ss (52); ds_ss (53); bd_ss (54);

 vp_ss (55); mr_ss (56); fw_ss (57);

  ps_ss (58); cd_ss (59); ss_ss (60);

OUTPUT:

 STDYX;

Appendix C

Five-factor Hierarchical Model

TITLE: Rachel Dissertation - Males Only

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

! Every variable in MODEL, does not include grouping variable

! GROUPING = male (0=Women 1=Men);

 MISSING = ALL (-999); ! Make sure to specify all missing values

useobs = male == 1;

ANALYSIS: ESTIMATOR = ML;

iterations=10000 ;

OUTPUT: !MODINDICES(0); ! Voodoo to improve model (list if p<.05 for df=1)

STDYX; ! Requests fully standardized solution (not shown here)

MODEL:

Factor1 by si_ss* vc_ss;

Factor1@1;

Factor2 by bd_ss* vp_ss;

Factor2@1;

Factor3 by mr_ss* fw_ss;

Factor3@1;

Factor4 by ds_ss* ps_ss;

Factor4@1;

Factor5 by cd_ss* ss_ss;

Factor5@1;

GenFac by Factor1* Factor2 Factor3 Factor4 Factor5;

GenFac@1;

TITLE: Rachel Dissertation – Females Only

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

! Every variable in MODEL, does not include grouping variable

GROUPING = male (0=women 1=men);

 MISSING = ALL (-999); ! Make sure to specify all missing values

useobs = male == 0;

ANALYSIS: ESTIMATOR = ML;

iterations=10000 ;

MODEL:

Factor1 by si_ss* vc_ss;

Factor1@1;

Factor2 by bd_ss* vp_ss;

Factor2@1;

Factor3 by mr_ss* fw_ss;

Factor3@1;

Factor4 by ds_ss* ps_ss;

Factor4@1;

Factor5 by cd_ss* ss_ss;

Factor5@1;

GenFac by Factor1* Factor2 Factor3 Factor4 Factor5;

GenFac@1;

OUTPUT:

STDYX;

TITLE: Five-factor model with males and females

DATA: FILE = rachelmplus.dat;

VARIABLE: NAMES = Male bd_ss si_ss mr_ss ds_ss cd_ss vc_ss fw_ss

vp_ss ps_ss ss_ss fsiq_ss vci_ss vsi_ss fri_ss wmi_ss psi_ss;

 USEVARIABLES = si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

! Every variable in MODEL, does not include grouping variable

 GROUPING = male (0=Women 1=Men);

 MISSING = ALL (-999); ! Make sure to specify all missing values

ANALYSIS: ESTIMATOR = ML;

iterations=10000 ;

OUTPUT: !MODINDICES(0); ! Voodoo to improve model (list if p<.05 for df=1)

 STDYX; ! Requests fully standardized solution (not shown here)

MODEL:

! Factor loadings all estimated

Factor1 by si_ss* (1);

Factor1 by vc_ss* (2);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (3);

Factor2 by vp_ss* (4);

Factor2@1;[Factor2@0];

Factor3 by mr_ss* (5);

Factor3 by fw_ss* (6);

Factor3@1;[Factor3@0];

Factor4 by ds_ss* (7);

Factor4 by ps_ss* (8);

Factor4@1;[Factor4@0];

Factor5 by cd_ss* (9);

Factor5 by ss_ss* (10);

Factor5@1;[Factor5@0];

! second order loadings;

GenFac by Factor1* (11);

GenFac by Factor2 (12);

GenFac by Factor3 (13);

GenFac by Factor4 (14);

GenFac by Factor5 (15);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

MODEL WOMEN:

! Factor loadings all estimated

Factor1 by si_ss* (91);

Factor1 by vc_ss* (92);

Factor1@1;[Factor1@0];

Factor2 by bd_ss* (93);

Factor2 by vp_ss* (94);

Factor2@1;[Factor2@0];

Factor3 by mr_ss* (95);

Factor3 by fw_ss* (96);

Factor3@1;[Factor3@0];

Factor4 by ds_ss* (97);

Factor4 by ps_ss* (98);

Factor4@1;[Factor4@0];

Factor5 by cd_ss* (99);

Factor5 by ss_ss* (910);

Factor5@1;[Factor5@0];

! second order loadings;

GenFac by Factor1* (911);

GenFac by Factor2 (912);

GenFac by Factor3 (913);

GenFac by Factor4 (914);

GenFac by Factor5 (915);

GenFac@1;[GenFac@0];

! item intercepts;

[si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss];

! Item variances

si_ss vc_ss ds_ss bd_ss vp_ss mr_ss fw_ss

 ps_ss cd_ss ss_ss;

OUTPUT:

STDYX;