An Empirical Investigation Comparing the WISC-III and WISC-IV: Implications for the Atkins Criterion

Author: Gordon E. Taub, University of Central Florida.
Email: Gordon.Taub@ucf.edu.

Abstract

Forensic psychological professionals implicitly assume that scores obtained on a revised intelligence test are equivalent to scores obtained on its predecessor and that the new instrument provides the better measurement of intelligence. One prong in Atkins cases is the onset of intellectual deficiency (ID) prior to age 18. When divergent scores on the Wechsler Intelligence Scale for Children-III (WISC-III; Wechsler, 1991) and the Wechsler Intelligence Scale for Children-IV (WISC-IV; Wechsler, 2003a) are presented, the court must make a determination regarding which test provides the most accurate assessment of an inmate's intellectual functioning. The results from the present study found that the WISC-IV is the better measure of intelligence when the criteria are the test publisher's measurement and scoring model, comprehensive assessment of intelligence, and the number of items at the floor and ceiling of the instrument.

Keywords: (Atkins, Intelligence, WISC, Wechsler)

The first Wechsler scale designed to measure the intellectual functioning of children was published in 1949. The Wechsler scales measured intelligence along two dimensions, a Performance factor (PIQ) and a Verbal factor (VIQ). In 1991, the third revision of the instrument was published, the Wechsler Intelligence Scale for Children-III (WISC-III; Wechsler, 1991). It is worth noting that Dr. Wechsler died in 1981; however, the Psychological Corporation credited Wechsler as the author postmortem. The WISC-III continued the tradition of measuring intelligence via the PIQ and VIQ, and provided a Full-Scale Intelligence Quotient (FSIQ). The WISC-III also included the same twelve subtests as the two previous editions (Wechsler Intelligence Scale for Children; Wechsler, 1949; Wechsler Intelligence Scale for Children- Revised; Wechsler, 1974). Changes to the 1991 test included new norms, item revisions and additions, improved psychometric properties, and an improved overall appearance (Wechsler, 1991). A new supplemental subtest, Symbol Search, was also included in the revised instrument.

During development, the WISC-III's publisher conducted factor analyses using the instrument's standardization sample and identified four new clinical indices: Freedom from Distractibility, Perceptual Organization, Processing Speed, and Verbal Comprehension. With the publication of the WISC-III, practitioners had the option to calculate scores using these four index (or factor) scores for clinical purposes; however, the clinical factor scores were independent of the core battery contributing to the instrument's PIQ, VIQ, and FSIQ measurement and scoring structure. Although the scores contrib-

Taub, G. E. (2014). An empirical investigation comparing the WISC-III and WISC-IV: Implications for the Atkins Criterion. *Open Access Journal of Forensic Psychology, 6,* 1-16.

uting to the calculation of the four clinical factors did not contribute to the calculation of FSIQ, research on the WISC-III indicated the four-factor measurement and scoring model provided a better fit to the instrument's standardization data when compared to the two-factor VIQ/PIQ measurement and scoring model (Taub, 2001).

The WISC-III was replaced by the Wechsler Intelligence Scale for Children- IV (WISC-IV; Wechsler, 2003a) in 2003. The WISC-IV represents the scale's first departure from the PIQ, VIQ, FSIQ measurement and scoring model. Prior to the publication of the WISC-IV, the measurement of intelligence on the Wechsler scales was *not* based on a theory of intelligence. Rather, Wechsler incorporated elements of abilities he thought were important into his tests (Wechsler, 2003b). Although Wechsler considered the subtests included within his instruments to be different *measures* of intelligence, their grouping into the PIQ and VIQ dichotomy remained unchanged for over 54 years—from the publication of the WISC in 1949 until the publication of the WISC-IV in 2003 (Wechsler 2003a). The revised measurement and scoring model of the WISC-IV represents an advancement in research-based framework to measure intelligence, "the new framework of the [WISC-IV] is based on theory and supported by clinical research and factor-analytic results" (p. 5; Wechsler, 2003b). One of the key theories contributing to the four-factor model is Carroll's "Three Stratum Theory" of intelligence (Carroll, 1993; Wechsler, 2003b).

WISC-IV Revision Goals

The WISC-IV's publisher identifies several key revision goals for the instrument. These include improving the instrument's theoretical foundation, increased developmental appropriateness, and improved psychometric properties (Wechsler, 2003b). The revision goal to improve the instrument's theoretical properties reflects the need to incorporate contemporary theories of intelligence and factor analytic results (e. g., Carroll 1993; 1997) into the instrument's measurement and scoring model. Within the instrument's measurement model, fluid reasoning, processing speed, and working memory are identified as areas of improvement for the WISC-IV. To address the need for increased sensitivity at the floor and ceiling of the instrument, the publisher added easier and more difficult items to all retained subtests. According to the publisher, this makes the WISC-IV a "more accurate measure of cognitive functioning [compared to the WISC-III] for children performing at the extreme ranges of cognitive ability" (p. 11; Wechsler, 2003b).

Using exploratory and confirmatory factor analysis the publisher provides evidence to support the instrument's new theoretical measurement and scoring model. The WISC-IV's improved theoretical foundation is a departure from the traditional measurement of intelligence using the first-order VIQ and PIQ factors; the instrument uses a model with four first-order factors. The new model is based in part on contemporary intelligence theory, which indicates the presence of several first-order broad cognitive abilities that, in combination, contribute to overall intellectual performance (e. g., Carroll 1993; 1997). Using Carroll's three-stratum theory as a partial foundation for the instrument, the WISC-IV measures four first-order Indices or factors representing five broad abilities. These are: the Verbal Comprehension Index, a measure of acquired knowledge; the

Perceptual Reasoning Index, a measure of fluid reasoning and visual-spatial thinking; the Processing Speed Index, a measure of efficient cognitive processing; and the Working Memory Index, a measure of working memory (Keith, Fine, Taub, Reynolds, & Kranzler, 2006; Wechsler, 2003b).

The publication of a revised instrument is often assumed by practitioners to represent an overall improvement in the measurement of intelligence when compared to its predecessor. One way to provide evidence to support score interpretations from a new instrument is to test the invariance of the instrument's measurement and scoring model across the revised instrument's age range. In short, a test of invariance examines the degree to which an instrument measures the same constructs or factor structure across the instrument's age range. In their study examining the invariance of the WISC-IV, Keith et al. (2006) found the WISC-IV measures the same constructs across the instrument's age range; however, their findings indicate the measurement and scoring model provided by the test's author do not provide the best explanation of the instrument's factor structure. Their research indicates scores on the WISC-IV are best explained by a model containing five first-order factors in contrast to the instrument's current four-factor model (i.e., Perceptual Reasoning, Processing Speed, Verbal Comprehension and Working Memory). Their results indicate scores from the WISC-IV's standardization sample are best explained by a five-factor model that replaces the Perceptual Reasoning factor with two first-order factors, Fluid Reasoning and Visual-Spatial Thinking. This research was recently replicated and indicates that separating the Perceptual Reasoning Index into two distinct factors, Fluid Reasoning and Visual-Spatial, provides the best fit to the instrument's standardization data (Weiss, Keith, Zhu, & Chen, 2013).

Atkins Cases

With the Supreme Court's ruling in *Atkins v. Virginia* (2002) a ban on the execution of individuals identified as mentally retarded went into effect. This was because the court's ruling found the execution of individuals with intellectual disabilities was a violation of the Eighth Amendment's ban on cruel and unusual punishments. Under the Court's ruling, an individual with significantly sub-average intelligence may qualify for relief under *Atkins* and not be executed; but this is not automatic. There are three separate prongs under Atkins and each prong must be met for an inmate to receive relief under Atkins. Although the interpretation of the criteria varies across states, the first prong under Atkins is performance on a measure of intelligence that is approximately two or more standard deviations below the mean of the instrument. The second prong of Atkins is concurrent significant deficits in adaptive functioning and the third prong is onset of the intellectual disability prior to age 18.

Although the discussion of each prong is beyond the scope of this paper, it is important to note that measurement error affects an individual's FSIQ on a test of intelligence. This is important in the first prong of Atkins, which requires significantly sub-average intellectual functioning. Because measurement error affects each individual's score on an intelligence test, an observed FSIQ is considered an estimate of the person's True FSIQ. A True FSIQ is defined as an individual's FSIQ without measurement error. To account for measurement error, the publisher of the Wechsler scales recommends providing confidence intervals around the individual’s observed FSIQ. For example, an individual with an observed FSIQ of 73 on the WISC-IV has a confidence interval ranging from 69 to 79 at a 95% level of confidence. Thus, there is a 95% level of confidence that the individual’s True FSIQ is between 69 and 79. Applying the standard error measurement indicates that, in this example, the individual’s observed FSIQ is within the Borderline range; however, the individual’s True FSIQ may be within the ID range. Although the standard error of measurement is a widely accepted psychometric property of intellectual assessment, courts within many states do not recognize standard error of measurement in Atkins cases. The importance of standard error of measurement in Atkins cases is a topic of debate and will be addressed by the U.S. Supreme Court in *Hall v. Florida*.

In addition to the standard error of measurement, FSIQ may also be affected by an examinee’s effort on an intelligence test. Another potential influence is the Flynn Effect (FE; Flynn, 1984, 2012). The FE is defined as a mean increase of about three FSIQ points per decade within the general population (Weiss, 2010). Although the phenomenon of the FE is widely accepted, the reasons for the FE are enigmatic (Williams, 2013). Nevertheless, the FE is important in the practice of psychology (Kaufman & Weiss, 2010). In its simplest terms, a person who is administered an intelligence test ten years after its publication will have an FSIQ score that is approximately 3 FSIQ points higher than the FSIQ that would have been obtained ten years earlier, at the instrument’s publication date. In practical terms, an inmate who obtained an FSIQ of 71 on the WISC-III in 2001 would potentially have obtained a FSIQ of 68 when the instrument was first published in 1991. Similar to the standard error of measurement, many states do not recognize the FE in Atkins cases.

At the time of Atkins, the WISC- III was the most popular test of intelligence for children and the Wechsler Intelligence Scale for Adults III (Wechsler, 1997) was considered the gold standard to measure the intellectual functioning of adults. Since Atkins was decided in 2002 and the WISC-IV was published in 2004 it is possible that an inmate was administered the WISC-III and later administered the WISC-IV, both prior to age 18. It is not unusual to observe divergence in the FISQ score across two versions of the same instrument. Divergence in FSIQ really becomes an issue in Atkins cases when the FSIQ from one instrument is in the Borderline range (i.e., FSIQ of 71 to about 79) while the FSIQ obtained from another instrument is within the Intellectually Deficient (ID) range (i.e., 70 and below). When this occurs, it is necessary for the court to determine which test, the WISC-III or the WISC-IV, is the better measure of intelligence.

When courts do not recognize the standard error of measurement error and/or the FE, decisions regarding ID determinations under Atkins are based solely on obtained FSIQ scores. When scores across the Wechsler scales are divergent, courts are faced with two possibilities. The first possibility is that the inmate’s FSIQ score on the WISC-III is in the Borderline range whereas the FSIQ on the WISC-IV is in the ID range. The second alternative is that the inmate’s obtained FSIQ score on the WISC III is in the ID range; in contrast, the FSIQ score on the WISC-IV is in the Borderline range. When either of these scenarios occurs, it is the court's responsibility to identify which test, the WISC-III or the WISC-IV, provides the best measure of the inmate's intellectual function prior to the age of 18.

Method

Study participants include the WISC-III and WISC-IV standardization samples. A total of 2,200 participants are included in the WISC-III's standardization sample; they range in age from 6 to 16 years of age. The WISC-IV standardization sample ranges between 6 and 16 years of age and includes a total of 2,200 participants. The Technical Manual of the WISC-III (Wechsler, 1991) and WISC-IV (Wechsler, 2003b) provide a more in-depth description of each instrument's participants.

Design of the Study

The averaged covariance matrix derived from each instrument's standardization data is used as input data for all analyses within the present study. These matrices are analyzed using structural equation modeling (SEM) via the AMOS 7. 0 (Arbuckle, 2007) statistical program following the method of maximum-likelihood estimation. The input data used in the study only includes the scores from tests contributing to FSIQ; scores from supplemental tests are not included in the analyses. One reason for this limitation is that, in general, forensic psychological professionals interested in the presence or absence of ID generally only administer core subtests (i.e., the tests that contribute to the calculation of FSIQ). Four models are tested. In Model 1, the standardization data from the WISC-III is used to test this instrument's VIQ, PIQ, FSIQ measurement, and scoring model. In Model 2 a new model is tested. In Model 2, the WISC-III's *clinical* four-factor model is modified to include a FSIQ measurement and scoring model. This model is presented in Figure 1. The decision to create and test this model is based on previous research investigating the Wechsler Adult Intelligence Scale-Third Edition's (WAIS-III; Wechsler, 1997) *clinical* four-factor model, which is similar to the WISC-III's four-factor model. The findings from this research indicate that the WAIS-III's *clinical* four-factor model, when fit into an FSIQ measurement and scoring model, provides an improved fit to the WAIS-III's standardization data when compared to the publisher's VIQ, PIQ, FSIQ measurement and scoring model (Taub, 2001; Taub, McGrew, & Witta, 2004). Model 3, presented in Figure 2, investigates the fit of the WISC-IV's four-factor measurement and scoring model to scores from this instruments standardization sample. The final model, Model 4, simultaneously tests the invariance of scores across the WISC-III and WISC-IV. Thus, Model 4 tests the comparability or equivalence of scores across the two versions of the Wechsler scales as well as issues related to the FE.

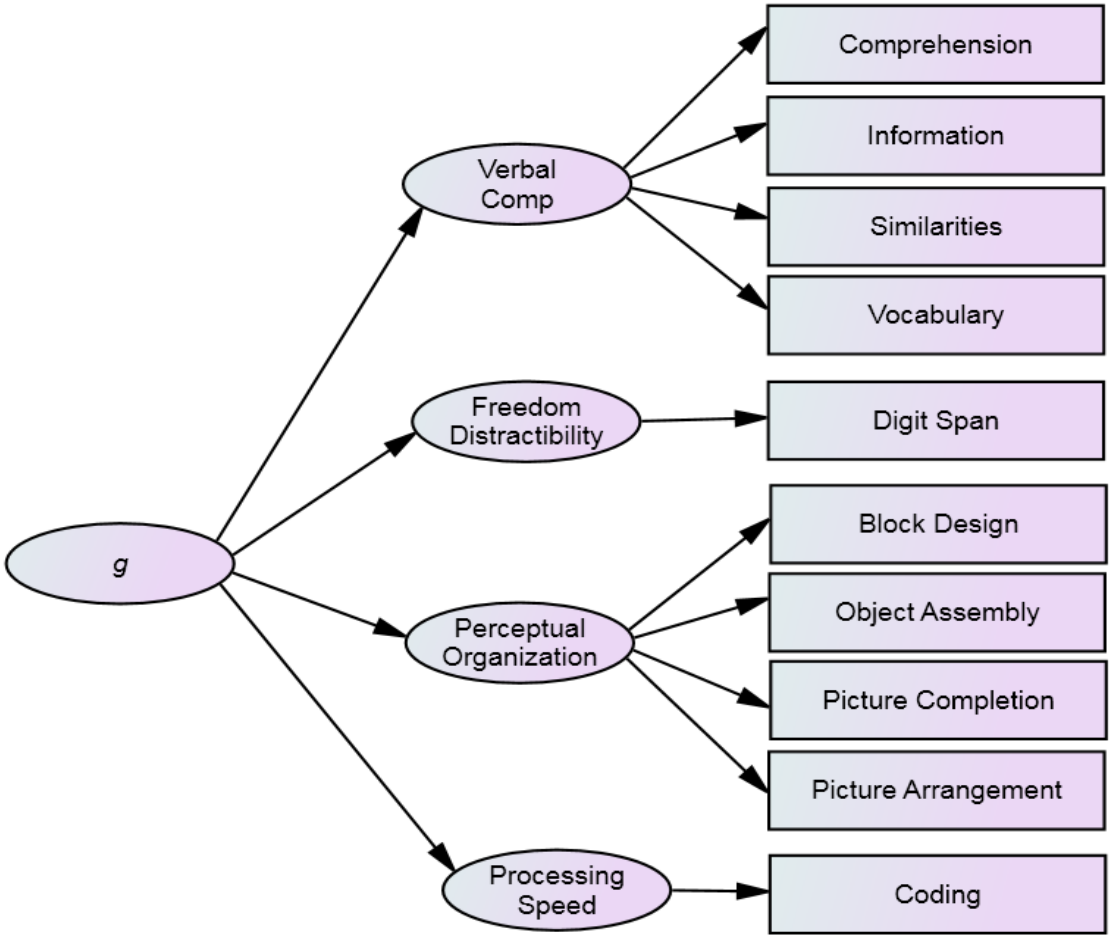

Figure 1. WISC-III *Clinical* Four-Factor Model

Taub, G. E. (2014). An empirical investigation comparing the WISC-III and WISC-IV: Implications for the Atkins Criterion. *Open Access Journal of Forensic Psychology, 6,* 1-16.

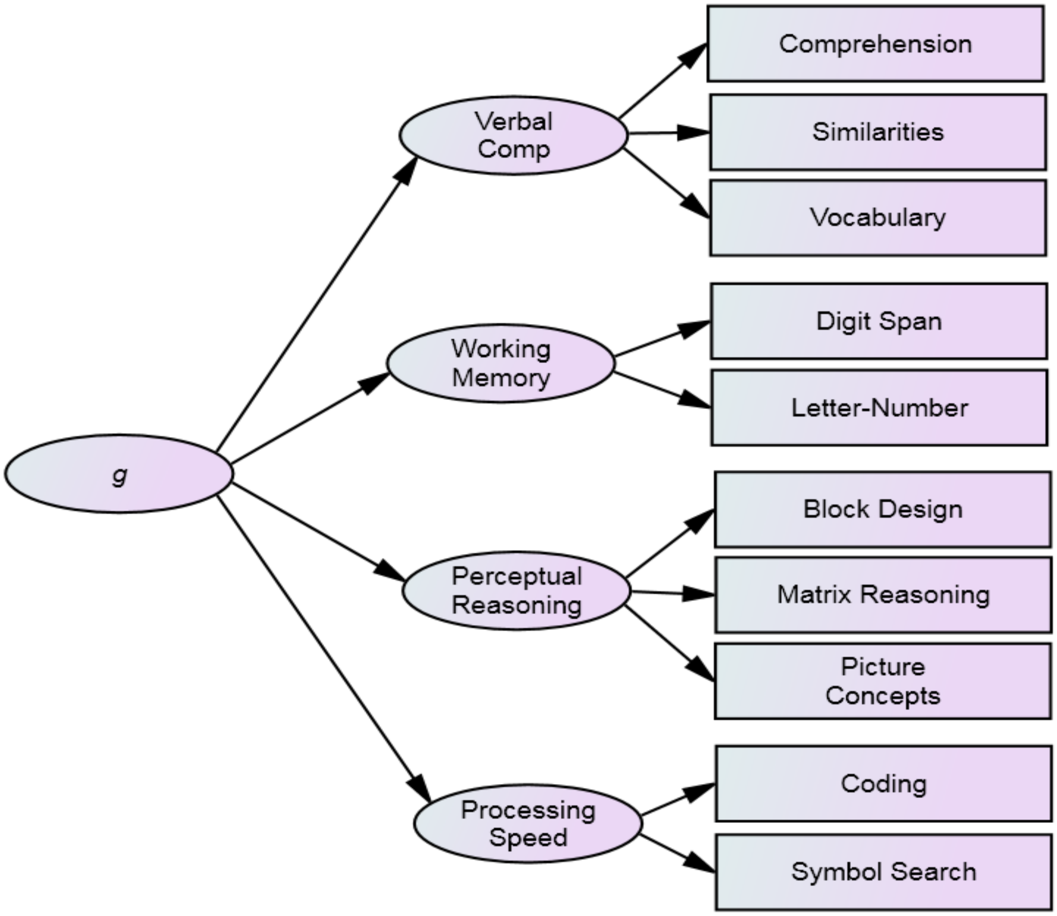

Figure 2. WISC-IV Four-Factor Model

The next analysis is an investigation of the portion of psychometric intelligence or *g* accounted for by scores from the ten subtests contributing to FSIQ. These analyses identify the portion of score variance that contributes to FSIQ and answers the question, *Which test, the WISC-III or the WISC-IV, provides the most internally consistent estimate of FSIQ?* First, coefficient omega hierarchical was used to account for all *g* variance (Reise, Scheines, Widaman, & Haviland, 2013). Second, coefficient omega was used to account for total test variance, and coefficient omega hierarchical was subtracted from coefficient omega to determine the extent of non-*g* variance contained in the FSIQ.

The final set of analyses investigates average variance extracted (AVE) and construct reliability (CR) of each instrument's first-order factors (e. g., fluid reasoning, processing speed). The AVE and CR analyses are used to investigate the level of support for the interpretation of each instrument's first-order factor scores as reliable and valid (AERA, APA, & NCME, 1999), thus providing data to compare the support for each instrument's first-order factors.

Use of Fit Indices

Several fit indices were utilized when evaluating the fit of the Wechsler scale's measurement and scoring models to the standardization data. These fit indices include the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI). Values for the CFI and TLI range from 0. 00 to 1. 00; values >.95 indicate an excellent fit and values > .90 indicate an adequate fit (Hu & Bentler, 1999). The Root Mean Square of Approximation (RMSEA) also was utilized when evaluating model fit; RMSEA values range from 0.00 to 1.00. A value of zero indicates a perfect fit; values equal to or less than .05 indicate a good fit, and values up to .10 indicate a mediocre fit (Byrne, 2010). When comparing two or more non-nested models, the Akaike Information Criteria (AIC) was used; the lower AIC indicates the best model. Although it is possible to compare model fit via the chi-square statistic, with the lower chi-square indicating the better fit, the AIC statistic provides the *best* index of change in fit across non-nested models (Keith et al., 2006).

As successive constraints are added across standardization samples in tests of invariance, the likelihood ratio test ($\Delta\chi^2$) is used to evaluate successively more restricted models (Keith, 2005; Taub et al, 2004). As increasingly restricted constraints are added across models, support for invariance is provided through statistically non-significant increases in the $\Delta\chi^2$. The RMSEA and CFI are used as secondary indices to evaluate tests of invariance because they perform reasonably well in simulation studies (Meade, Johnson, & Braddy, 2008).

Results

The present study investigates the recent editions of the Wechsler Intelligence Scales for Children. All analyses only include scores from tests contributing to the calculation of FSIQ; scores from supplemental tests are not included in the analyses. The reasons for only using scores from core subtests include 1) in general, psychological professionals only administer tests contributing to FSIQ when the outcome of interest is the presence or absence of ID and 2) the instruments' publisher conducted separate analyses with core subtests.

Model 1

The results from this analysis are presented in Table 1. The fit indices associated with Model 1 indicate the WISC-III's VIQ, PIQ, FSIQ measurement model provides a mediocre fit to the instrument's standardization data. The next analysis using WISC-III scores, Model 2, is based on prior research.

Table 1

Fit Statistics for Scoring Measurement Models

Phase/Model	χ^2 *(df)*	AIC	CFI	RMSEA	TLI
Phase 1					
Model 1	321 (34)	363.04	.97	.062	.96
Model 2	277.08 (33)	321.08	.975	.058	.967
Model 3	171.43 (31)	219.43	.984	.045	.977

Note. AIC= Akaike Information Criteria, CFI = Comparative Fit Index, TLI = Tucker Lewis Index, RMSEA = Root Mean Square Error of Approximation.

Model 2

Model 2 tests the fit of the WISC-III subtest scores to the instrument's *clinical* four-factor model and is displayed in Figure 1. When comparing two non-nested models, such as Model 1 to Model 2, the best model is generally identified as the one with the lowest AIC. An examination of the AIC between Model 1 and 2 indicates the AIC associated with the WISC-III's *clinical* four-factor model provides a better fit to the standardization data than the publisher's VPIQ measurement and scoring model—219.43 and 363.04, respectively.

Model 3

Model 3 tests the fit of the WISC-IV's standardization data to the instrument's four-factor FSIQ measurement and scoring model as presented in Figure 2. The results from this analysis, presented in Table 1, indicate that the WISC-IV's four-factor model provides a good fit to the data, as indicated by the RMSEA (.045) in Table 1 as well as the other fit indices.

Model 4

The final model, Model 4, simultaneously tests the invariance of scores from the WISC-III and WISC-IV to determine if the observed scores across both instruments are equivalent. The results from this analysis found the model to be unidentified. One limitation within Model 4 is that the WISC-III and WISC-IV share only five core subtests in common. Because of this limitation, it is not possible to test the invariance of a four-factor measurement and scoring model. Although it may be possible to test the invariance of a two-factor model (i.e., VPIQ), it was not tested due to the model's limited theoretical support. The limitations for support of the VPIQ model include: 1) results from model comparisons indicate that the WISC-III's *clinical* four-factor measurement and scoring model provides the best fit to the WISC-III's standardization data and 2) the WISC-IV is based on a four-factor measurement and scoring model. Because of these limitations, results from a test of invariance based on two factor first-order VPIQ factors are not reported herein.

FSIQ g-Loadings

The results indicate that both the WISC-III and WISC-IV have nearly identical *g* loadings; coefficient omega hierarchical was .8437 and .8411 respectively.

First-Order Factors

The next set of analyses focuses on the extent to which the WISC-IV's publisher reached the goal of improving the instrument's theoretical foundation via improved measurement of fluid reasoning, processing speed, and working memory. This analysis examined common variance via the Average Variance Extracted (AVE) and internal consistency via construct reliability (CR) of each instrument's first-order factor scores. This analysis is conducted using a model wherein all first-order factor scores are correlated. Results from the WISC-III AVE and CR analyses serve as a baseline to identify the change in AVE and CR estimates to the WISC-IV. The AVE is calculated by obtaining the sum of squared factor loadings then dividing the sum by itself plus the sum of standard error variances. The formula for CR is similar; however, factor loadings are summed prior to being squared. Values greater than .5 for AVE and .7 for CR are considered minimally acceptable. The results of the AVE and CR estimates for the WISC-III and WISC-IV are presented in Table 2. AVE and CR estimates were not available for the WISC-III's Freedom from Distractibility and Processing Speed factors because only one subtest contributes to the calculation of these factors when the outcome of interest is scores on subtests contributing to the calculation of FSIQ. Therefore, when only administering tests contributing to the calculation of FSIQ, the WISC-IV clearly provides a better measure of Processing Speed and Working Memory relative to the WISC-III. The results also indicate that the AVE is higher for the WISC-IV's Verbal Comprehension and Perceptual Reasoning factors relative to analogous factors on the WISC-III. The CR estimate is slightly higher on the WISC-III's Verbal Comprehension factor when compared to the WISC-IV (i.e., a difference of .008) and on the Perceptual Reasoning factor (a difference of .045).

Table 2

AVE and CR for First-Order Factors

Factor		WISC-III	WISC-IV
Verbal Comprehension	AVE	.639	.685
	CR	.875	.867
Perceptual Organization/	AVE	.473	.482
Reasoning	CR	.780	.735
Freedom from Distractability/	AVE	-	.489
Working Memory	CR	-	.655
Processing Speed	AVE	-	538
	CR	-	.697

Note. AVE = average variance extracted, CR = construct reliability, NC = Not calculated.

Discussion

This study investigates the measurement and scoring models of the WISC-III and WISC-IV. The purpose of the study is to identify which instrument is better, the WISC-III or the WISC-IV. This study first examines the factor structure of the Wechsler scales in an effort to identify how well each instrument fits the publisher's measurement and scoring model. This addressed the following questions: *Do the WISC-III and WISC-IV measure intelligence as the publisher intended and is the publisher's scoring model the best way to account for performance on each instrument?* Next was an investigation of the invariance of scores across versions of the Wechsler scales. This was done to answer the question: *Are scores across the WISC-III and WISC-IV equivalent (e.g., is an FSIQ score of 78 on the WISC-III directly comparable to an FSIQ score of 78 on the WISC-IV?).* The next analysis compared the portion of variance accounted for by psychometric *g* across both instruments. The final analysis investigated how well the publisher met its WISC-IV revision goals. This answered the question: *Does the WISC-IV provide better measurement of fluid reasoning, processing speed, and working memory when compared to the WISC-III?*

Models 1 and 2

The first models tested, Models 1 and 2, investigate the fit of scores from the WISC-III standardization data to two different measurement and scoring models. First, Model 1 tested the fit of scores to the instrument's VIQ, PIQ, FSIQ measurement and scoring model. Next, Model 2 tested the fit of scores from the standardization data to the instrument's *clinical* four-factor model, as presented in Figure 1. When comparing non-nested models, such as Models 1 and 2, the AIC is the preferred statistic; the model with the lowest AIC provides the best fit to the data. The AIC of Model 2 is lower than Model 1, 321.08 compared to 363.04, thus the AIC identifies Model 2 as the best fitting model. Other fit indices presented in Table 1 also suggest that Model 2 is the best fitting model. Thus, scores on the core subtests of the WISC-III are best interpreted via the *clinical* four-factor model, which includes four first-order factors in contrast to the instrument's traditional VPIQ measurement and scoring model consisting of two first-order factors.

Models 3 and 4

Model 3, presented in Figure 2, is a test of the fit of the WISC-IV's standardization data to the instrument's four first-order factor model; this hierarchical model includes FSIQ. As presented in Table 1, the fit indices associated with Model 3 all indicate the WISC-IV's four-factor model provides an adequate fit to the scores from the instrument's core subtests. The next model, Model 4 investigates the equivalence of scores across instruments. Because only five core subtests are shared across the WISC-III and WISC-IV, the model lacks construct representation and is underidentified, thus no results are available from this analysis.

FSIQ g-Loadings

Analyses of coefficient omega hierarchical were calculated for both the WISC-III and the WISC-IV. Results from these analyses indicate FISQ subtest score variance associated with psychometric *g* was nearly identical across both Wechsler scales. Therefore, the WISC-III and WISC-IV account for a similar and adequate portion of *g* variance within FSIQ subtest scores.

First-Order Factors

Results from the AVE and CR estimates provide support for the attainment of the WISC-IV's revision goals related to improved measures of Fluid Reasoning, Processing Speed, and Working Memory. The WISC-IV appears to have better indicators of Verbal Comprehension, Perceptual Reasoning, Processing Speed, and Working Memory relative to WISC-III indicators of analogous factors based on higher AVE estimates. Additionally, relative to the WISC-III, CR estimates are lower for the WISC-IV factors Verbal Comprehension (a difference of .008) and Perceptual Reasoning (a difference of .083). The lower reliability for the WISC-IV Perceptual Reasoning and Verbal Comprehension factors likely reflects the inclusion of fewer subtests to measure these constructs (i.e., four subtests for the WISC-III relative to three subtests per factor for the WISC-IV. Additionally, previous research suggests Perceptual Reasoning factor on the WISC-IV is a mixed measure of Fluid Reasoning and Visual-Spatial Thinking (Keith et al., 2006), which supports the conclusion that measurement of Fluid Reasoning was improved in addition to Processing Speed and Working Memory.

Implications for Practitioners

Which test is better, the WISC-III or the WISC-IV, may be addressed in several ways. The first way is to investigate how well each instrument fits the publisher's measurement and scoring model. The results from this study indicate that the WISC-IV scores provide an adequate fit to the publisher's measurement and scoring model, although there is room for improvement. In comparison, when only using scores from subtests contributing to the calculation of FSIQ the WISC-III does not provide an adequate fit to the publisher's measurement and scoring model, so the WISC-IV is the better instrument in this area.

Most intelligence tests published within the past ten years are consistent with CHC theory (Keith & Reynolds, 2010). CHC theory identifies seven broad areas of intelligence. A comprehensive test of intelligence is considered one that measures many CHC broad abilities. Although measuring more abilities is thought to provide a better measure of *g*, both instruments account for a similar portion of *g* variance, there was no difference across instruments.

The publisher of the WISC-III indicates that the WISC-IV was designed to provide a better measure of the CHC broad abilities *fluid reasoning*, *processing speed*, and *working memory* (when compared to the WISC-III). It is important to note that the inclusion of scores from the supplemental tests in this study's analyses may yield different results. However, the focus of this study is the investigation of the differences across instruments when only administering subtests contributing to the calculation of FSIQ. The results from the investigation in the broad abilities measured by the Wechsler scales found that the WISC-IV provides a better measure of fluid reasoning, processing speed, and working memory relative to the WISC-III, when the outcome of interest is scores from subtests contributing to FSIQ.

Additionally, the inclusion of the Matrix Reasoning subtest supports the WISC-IV as providing the better measure of Fluid Reasoning. Thus the WISC-IV provides a broader and more comprehensive measure of factors associated with intelligence when compared to the WISC-III. As previously discussed, in revision, several items were added to the floor (e. g. , ID and below) and ceiling (e.g. high Average and gifted) of the WISC-IV. These questions were added to provide a better measure of intellectual function at the extreme ranges of the instrument. Thus, individuals scoring in the lower range of FSIQ will have more questions to answer, which provides a more reliable score. Similarly, more difficult questions were added to provide a more reliable score in the higher ranges of FSIQ. In Atkins cases wherein individuals are scoring at the lower end of the instrument, the addition of more questions is believed to result in a more reliable FSIQ.

Taken together, when comparing only scores contributing to FSIQ, both instruments provide similar measures of psychometric *g*. The WISC-IV provides a better measurement of intelligence when the criterion is alignment with the publisher's measurement and scoring model and comprehensively measuring intelligence. The WISC-IV's publisher also states that this instrument provides improved reliability at the extreme ranges of the instrument (i.e., ID and gifted). The more comprehensive measurement of intelligence by the WISC-IV may also be viewed as a technological advancement. This is because the Wechsler scales did not adequately measure fluid reasoning and/or working memory for over 50 years, until the publication of the WISC-IV. The WISC-IV is less reliant on abilities associated with Perceptual Organization than the WISC-III, which provides the opportunity to more comprehensively measure Fluid Reasoning, Processing Speed, and Working Memory. Using these criterions, the WISC-IV allows assessment of more abilities and allows users to derive more scores, which results in a more comprehensive assessment of intelligence. The measurement model contains multiple factors, which would be operationalized by deriving multiple scores. The WISC-IV provides a more reliable, valid, and comprehensive measure of intelligence, which is also more consistent with the publisher's theoretical model to measure intelligence than the WISC-III. Therefore when applying the Atkins criterion and deciding which instrument is better, the WISC-III or the WISC-IV, the results from the current study indicate the WISC-IV provides a more reliable, valid, and comprehensive measure of intelligence compared to the WISC-III.

References

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

Arbuckle, J. L. (2007). *Amos 7.0* [Computer software]. Chicago, IL: Smallwaters.

Atkins v. Virginia, 536 U.S. 304 (2002).

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L.

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

Flynn, J. R. (2012). *Are we getting smarter? Rising IQ in the twenty-first century*. New York, NY: Cambridge University Press.

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

Kaufman, A. S., & Weiss, L. G. (2010). Flynn effect [Special issue]. *Journal of Psychoeducational Assessment, 28*(5).

Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds*.), Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 581-614). New York: Guilford Press.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47,* 635–650.

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* (1), 108-127.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative indices in tests of measurement invariance. *Journal of Applied Psychology, 93,* 568-592.

Reise, S, P., Scheines, R., Widaman, W. F., & Haviland, M. G. (2013). Multi-dimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5-26.

Taub, G. E. (2001). A confirmatory analysis of the Wechsler Adult Intelligence Scale-Third Edition: Is the Verbal/Performance Discrepancy Justified? *Practical Assessment, Research, & Evaluation*. Retrieved from http:/pareonline.net/getvn.asp?v=7&n=22

Taub, G. E., McGrew, K. S. & Witta, E. L. (2004). A confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale-Third Edition. *Psychological Assessment*, 16 (1), 85-89. Doi: 10.1037/1040-3590.

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

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale-Third Edition*. San Antonio, TX: The Psychological Corporation. Wechsler, D. (1997b). *WAIS-III WMS-III Technical Manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children-Fourth Edition*. San Antonio, TX: Pearson Assessment.

Wechsler, D. (2003b). *WISC-IV Technical and Interpretive Manual.* San Antonio, TX: Pearson Assessment.

Weiss, L. W. (2010). Considerations on the Flynn Effect. *Journal of Psychoeducational Assessment,* 28 (5), 482-493. Doi: 10.177/0734282910377372.

Weiss, L. W., Keith, T. Z., Ahu, J., & Chen, H. (2013). WISC-IV and clinical validation fo the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment, 31*(2), 114-131. Doi: 10.1177/073428291347803.

Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence*, *41*(6), 753-764. doi:10.1016/j.intell.2013.04.010