Journal of Forensic Psychology Practice, 13:27–48, 2013

ISSN: 1522-8932 print/1522-9092 online
DOI: 10.1080/15228932.2013.746913

Routledge
Taylor & Francis Group

Matters of Consequence: An Empirical Investigation of the WAIS-III and WAIS-IV and Implications for Addressing the Atkins Intelligence Criterion

GORDON E. TAUB, PhD
School Psychology Program, University of Central Florida, Orlando, Florida

NICHOLAS BENSON, PhD
Counseling and Psychology in Education, The University of South Dakota, Vermillion, South Dakota

"Which test provides the better measurement of intelligence, the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III) or the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV)?" is an important question to professional psychologists; however, it has become a critical issue in Atkins cases wherein courts are often presented with divergent Full-Scale IQ (FSIQ) scores on the WAIS-III and WAIS-IV. In these instances, courts are required to render a decision stating which test provided the better measure of an inmate's intellectual functioning. This study employed structural equation modeling to empirically determine which instrument; the WAIS-III or the WAIS-IV, provides the better measure of intelligence via the FSIQ score. Consistent with the publisher's representation of intellectual functioning, the results from this study indicate the WAIS-IV provides superior measurement, scoring, and structural models to measure FSIQ when compared to the WAIS-III.

KEYWORDS Atkins, intelligence, intelligence quotient, expert testimony, Wechsler Adult Intelligence Scale

Address correspondence to Gordon E. Taub, School Psychology Program, University of Central Florida, 4000 Central Florida Blvd., Suite CED 123K, Orlando, FL 32816. E-mail: gtaub@ucf.edu

HISTORICAL OVERVIEW

The first Wechsler scale, the Wechsler-Belleview Intelligence Scale (WBIS; Wechsler, 1939) was published in 1939. The WBIS divided the measurement of intelligence into two factors: Verbal and Performance. These two factors were combined, and their sum was transformed into one's full-scale intelligence quotient (FSIQ).

The Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) was created as a revision to the WBIS in 1955. The WAIS included revisions to the instrument's administration, items, and scoring. The WAIS retained the 11 subtests included in the WBIS. The WAIS was followed by the publication of the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981). The revision to the WAIS-R included administration, item, and scoring changes. All 11 subtests from the WBIS and the WAIS were retained in the WAIS-R.

The WAIS-R was replaced by the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997a). The WAIS-III retained the same subtests from the WAIS-R; however, three new subtests were added. The subtest Matrices was the only new subtest contributing to the calculation of an individual's FSIQ; the other two subtests were supplemental. The WAIS-III was scored in a manner identical to that of its predecessors using the traditional Verbal IQ, Performance IQ, and FSIQ scoring structure. There was one significant addition to the WAIS-III: a new four-factor *clinical* model, which was independent from the instrument's FSIQ measurement and scoring model. Although the research has demonstrated the WAIS-III's measurement model is invariant across the instrument's standardization sample, the new four-factor *clinical* model provided a better fit to the WAIS-III's standardization data when compared to the FSIQ measurement and scoring model (i.e., Verbal/Performance scoring model; Taub, 2001; Taub, McGrew, & Witta, 2004).

The FSIQ scoring model of the Wechsler scales, via the Verbal/Performance scoring method, was present in the Wechsler series from the publication of the WBIS in 1931 through the publication of the WAIS-III in 1997 (Wechsler, 1997b). Although theories of intelligence were published in empirical literature of the time, the Wechsler scales retained the verbal and performance scoring dichotomy to measure intelligence for over 60 years. Nevertheless, Wechsler and the test's publisher were aware of advances in intelligence theory and factor analytic research that identified important cognitive constructs beyond verbal and performance abilities (Wechsler, 2008a). The WAIS-III was replaced by the Wechsler Adult Intelligence Scale (WAIS-IV; Wechsler, 2008a), which is the *first* Wechsler scale, in the adult series, developed using a "new framework [that] is based on current intelligence theory and supported by clinical research and factor-analytic results" (Wechsler, 2008b, p. 8).

THE WAIS-IV

Revision Goals

One of the revision goals of the WAIS-IV was to provide a theoretical foundation for the measurement of intelligence. This included the incorporation of contemporary research and factor analytic results into the scoring and measurement model of the WAIS-IV (Wechsler, 2008b). One of the main theories incorporated into the WAIS-IV is Carroll's Three Stratum Theory (e.g., Carroll, 1993, 1997), which in part served as a theoretical blueprint for the instrument's revision. Based on their theoretical blueprint, the publisher identifies the measurement of fluid reasoning, processing speed, and working memory as areas of weakness of the WAIS-III and areas targeted for revision within the WAIS-IV (Wechsler, 2008b).

Another revision goal of the WAIS-IV addressed improved psychometric properties including updated norms, improved reliability and validity, and extended floors and ceilings. The updating of norms was based partially on research indicating that older norms produce inflated scores on intelligence tests (Flynn, 1984; Weiss, 2010). Therefore, this revision goal is not unique to the WAIS-IV and cannot be viewed as a weakness of the WAIS-III.

Revisions

The WAIS-IV's reliability and validity were examined through concurrent validity and reliability studies as well as confirmatory factor-analytic studies using first the core subtest contributing to FSIQ, followed by analyses including core and supplemental subtests.

The WAIS-IV provides improved measurement at the floor and ceiling of the instrument "to ensure an adequate range of scores to represent a broad range of cognitive ability, from extremely low (i.e., $40 < \text{FSIQ} < 69$) to very superior (i.e., $130 < \text{FSIQ} < 160$)" (Wechsler, 2008a, p. 23). Therefore, the publisher indicates that the inclusion of additional items on the WAIS-IV resulted in an improvement in reliability and validity of FSIQ for individuals scoring at or greater than two standard deviations above and two standard deviations below the instrument's mean.

The improvement in the WAIS-IV's (1) measurement of fluid reasoning, processing speed, and working memory, (2) improved reliability and validity especially at the extreme ranges of the instrument, and (3) extended floor and ceilings required extensive revision to the nine subtests retained from the WAIS-III in the WAIS-IV revision. A summary of the major changes to the subtests contributing to FSIQ and the first-order factors associated with each subtest is presented in Table 1. All of the subtests presented in Table 1 went through major revisions, which include item addition and elimination, removal/addition of item types, additional subtest activities, modification of

TABLE 1 Changes to the WAIS-IV Subtests

First-Order Factor	Subtest and Change
Verbal Comprehension	Information: 11 of 26 items are new and scoring criteria was changed Similarities: 12 of 18 items are new and corrective feedback is now provided Vocabulary: Three new picture items, six new verbal items and new scoring
Perceptual Reasoning	Block Design: Four of fourteen items are new, instructions were shortened, bonus points were eliminated, BDN process score was added Figure Weights: New test Matrix Reasoning: Item types reduced to two from four, 14 of 26 items are new
Working Memory	Digit Span: Five trials of digit span forward and nine trials of digit span backward are new, two new sample items, the addition of Digit Span Sequencing as a third test was added as were six process scores. Arithmetic: 11 of 12 retained items were revised, 9 new items were added, presentation blocks were eliminated, and picture items were added
Processing Speed	Coding: Two symbols were retained but revised, four new symbols were added, samples were increased from four to six, each number now appears twice in a row and total items were increased Symbol Search: Examinee must now mark the symbol or the NO box, instructions were simplified, symbols were enlarged

directions, and changes in scoring and/or item presentation. Interestingly, all nine of the retained subtests in the WAIS-IV have changes in recording and scoring as well as the inclusion of new items (Wechsler, 2008b). Previous research indicates that changing just the directions of a subtest can change the properties of the underlying measured construct to the extent that a comparison of scores between a revised subtest and the previous version would be like comparing apples and oranges (Kaufman, 2010). Taken together, it may appear that although subtest names remained the same, the constructs measured by each subtest may not be equivalent across the WAIS-III to WAIS-IV revision.

Theoretical Foundation

The improved theoretical foundation of the WAIS-IV resulted in the publisher's departure from the first-order Verbal IQ Performance IQ (VPIQ) Index scoring dichotomy to a new four-factor scoring structure. Thus, the WAIS-IV represents a new era in the Wechsler adult series. This four-factor model, based in part on Carroll's Three Stratum Theory of intelligence,

consists of a Processing Speed Index, which provides a measure of processing speed; a Perceptual Reasoning Index, providing a measure of fluid reasoning and visual-spatial abilities (Benson, Hulac, & Kranzler, 2010); and a Working Memory Index, a measure of working memory. Assumptions made by psychologists include that all revised intelligence tests provide a better measure of intelligence when compared to the instrument it replaced as well as that the publisher successfully met all of the instrument's revision goals within the new instrument. Consequently, it is assumed that the WAIS-IV's FSIQ score represents an improved or better FSIQ score when compared to the WAIS-III's FSIQ score; however, research supporting this assumption is lacking. Although the question "Does the WAIS-III or the WAIS-IV provide the most valid FSIQ score when the criterion of reference is the test publisher's representation and measurement of intellectual functioning?" is an important question to answer, it is of particular importance to forensic psychologists and legal professionals working in the area of Atkins cases.

ATKINS CASES

On June 20, 2002, the United States Supreme Court ruled the execution of individuals with intellectual disabilities is a violation of the Eighth Amendment's ban on cruel and unusual punishment. This ruling by the Supreme Court in *Atkins v. Virginia* (2002) resulted in a ban on the execution of death row inmates identified as *mentally retarded*. After the Court ruled in favor of Atkins, many inmates sentenced to death for capital crimes filed lawsuits to stop their execution by claiming the death penalty constitutes cruel and unusual punishment under the Eighth Amendment. These lawsuits are commonly referred to as *Atkins* cases.

At the time the Supreme Court ruled in favor of Atkins, the WAIS-III was the current version of the venerable Wechsler scales of intelligence. In its ruling, the Supreme Court considered the WAIS-III the "standard instrument in the United States for assessing intellectual functioning" (*Atkins v. Virginia*, 2002, p. 2245). When intelligence tests are revised, best-practice guidelines suggest practitioners should adopt the new instrument within 1 year of its publication. This means by 2009, most practitioners using the Wechsler adult series were using the WAIS-IV. However, since *Atkins* was decided in 2002, many inmates suspected of being intellectually deficient were assessed using the WAIS-III, not the WAIS-IV. In Atkins cases, the FSIQ score derived from the Wechsler scales may be viewed as the preferred indicator of intellectual functioning.

In many cases, the inmate's obtained FSIQ score on the WAIS-III is within the Borderline range, in contrast to the Intellectually Deficient (ID) range. The ID range is defined as two or more standard deviations below

the instrument's mean. On all Wechsler scales, this is an FSIQ score of 70 or below. The Borderline range includes FSIQ standard scores ranging from a low of 71 to a high of about 80. Individuals with FSIQ scores above 71 generally are not considered ID and will not receive relief under the Atkins standard.

An individual with an FSIQ of 70 or below *may* qualify under the Atkins standard. The reason why a FSIQ in the ID range does not automatically qualify one for relief under the Atkins standard is because ID is considered a pervasive developmental delay. To identify an individual as having an intellectual disability (formerly referred to as mental retardation), states generally require an inmate to meet three prongs. The first prong is a score reflecting overall performance on a measure of intelligence that is at least two standard deviations below the normative mean of the instrument. The second is onset of intellectual disability prior to age 18, and the third is displaying adaptive functioning that is sub-average or commensurate with expectations for individuals with an intellectual disability. These criteria vary across states, and a discussion of the differences between the prongs across states is beyond the scope of this study.

Practitioners familiar with basic psychometrics understand that measurement error affects all observed scores. Consequently, an individual's obtained FSIQ is considered *an estimate* of his or her True FSIQ score. The True FSIQ score is easily understood as the individual's FSIQ without measurement error. Test publishers provide confidence intervals around obtained FSIQ scores to account for measurement error. For example, an obtained FSIQ score of 72 on the WAIS-III has a confidence interval ranging between 67 and 77. This means that there is a 95% level of confidence that an individual who obtains an FSIQ score of 72 on the WAIS-III has a True FSIQ score between 67 and 77. One can easily see that a FSIQ of 72 is within the borderline range; however, when applying the confidence interval to account for measurement error, the individual's True score may be within the ID range. However, in many states courts do not recognize the standard error of measurement in Atkins cases.

Another factor that impacts an inmate's FSIQ is the Flynn effect (Flynn, 2007). The Flynn effect is commonly known among psychologists as an increase in FSIQ of about three points per decade. This effect has been demonstrated across a large number of studies, cultures, and tests (Weiss, 2010). The impact of the Flynn effect on FSIQ means that an inmate who was evaluated in 2007 on the WAIS-III will have an FSIQ score inflated by about three points. This is of particular importance in Atkins cases where there is a bright line of two standard deviations below the mean (i.e., FSIQ of 70) for the identification of ID. In other words, an inmate who was administered the WAIS-III in 2007 and received an FSIQ of 72 would potentially have scored 69 in 1998.

In addition to not recognizing the standard error of measurement, many states do not recognize the Flynn effect. In these states, the difference between one, two, or even three points could literally mean the difference between life and death. Therefore, it is not unusual for an inmate who has an FSIQ within the Borderline range on the WAIS-III to be administered the WAIS-IV, nor is it unusual for an inmate with an FSIQ in the ID range on the WAIS-III to be reevaluated using the WAIS-IV.

In instances when an inmate was administered the WAIS-III prior to 2009 and the state or defense requests a current FSIQ score, and the WAIS-IV is administered, two controversial outcomes are possible. The first outcome is an inmate's FSIQ score on the WAIS-III is within the borderline range and on revaluation obtains a WAIS-IV FSIQ in the ID range (i.e., 70 or below). Conversely, an inmate with an FSIQ in the ID range on the WAIS-III on a reevaluation may receive an FSIQ in the Borderline range on WAIS-IV. When either scenario occurs, the courts must decide which FSIQ score should be used to determine the first prong in Atkins: the WAIS-III or the WAIS-IV? The purpose of this study is to answer the question "Which test, the WAIS-III or the WAIS-IV, provides the most valid FSIQ score when the criterion of reference is the test publisher's representation and measurement of intellectual functioning?"

METHOD

Participants

This study's participants included the WAIS-III and WAIS-IV standardization samples. The WAIS-III was standardized using 2,450 participants ranging from 16 to 89 years of age. The WAIS-IV included 2,200 participants ranging from 16 years of age to more than 90 years of age. Both instruments divide the normative sample into 13 distinct age levels. The Technical Manual of the WAIS-III (Wechsler, 1997b) and WAIS-IV (Wechsler, 2008b) provides an in-depth description of each instrument's standardization sample.

Analyses

All analyses in this study were conducted using confirmatory factor analysis via the AMOS 7.0 (Arbuckle, 2007) statistical program. All analyses were conducted using the averaged covariance matrix derived from each instrument's standardization data, following the method of maximum-likelihood estimation via structural equation modeling. There were a total of four sets of analyses employed in the study. Because most psychologists administer only the subtests required to obtain an individual's FSIQ score, we limited our data input to the subtests contributing to the calculation of the FSIQ, from both Wechsler scales, in all analyses. It is worth noting the WAIS-IV's publisher

initially limited its confirmatory factor analyses to the core subtest contributing to the calculation of FSIQ (Wechsler, 2008b); the present study extends this area of research. The first analysis tested three models. The first analyses investigated the fit of the standardization data from each Wechsler scale to each respective instrument's measurement and scoring model. These analyses addressed the question "Which instrument provides the best fit to the publisher's measurement and scoring model?" Model 1 tested the measurement model derived from the WAIS-III's measurement and scoring model. Model 2 tested the WAIS-IV's measurement and scoring model; this model is similar to the factor model in Figure 5.1 of the WAIS-IV's Technical Manual (TM; Wechsler, 2008b, p. 67). In Model 3, the WAIS-III's *clinical* four-factor measurement model was modified to include an FSIQ measurement and scoring model (Taub, 2001; Wechsler, 1997a). The second analysis fit both the WAIS-III and WAIS-IV tests into a measurement and scoring model based on the Cattell-Horn-Carroll model (Benson et al., 2010; McGrew, 2009).

The third analysis investigated the measurement invariance between the WAIS-III and WAIS-IV standardization samples simultaneously. The simultaneous testing of invariance addresses several issues. First, the finding of invariance across instruments provides support for the equivalence of scores across versions. Score equivalence means it is possible to directly compare scores between the WAIS-III and WAIS-IV; thus, scores across the two versions represent a comparison of *apples to apples*. Additionally, the finding of equivalence provides support for the Flynn effect, meaning that differences in scores are due to generational improvements in intelligence in contrast to improvements in methodology and/or psychometrics (Beaujean & Osterlind, 2008; Brand, 1996; Rodgers, 1999; Wicherts et al., 2004); changes in item development including administration, presentation, and scoring (Kaufman, 2010); or theoretical differences across versions in the measurement of intelligence (McGrew, 2010). In other words, although both instruments include subtests with the same name, a test of invariance answers the question "Do subtests with the same name measure identical constructs across instruments?" The WAIS-III requires the administration of 11 subtests to calculate FSIQ; 10 subtests contribute to the calculation of FSIQ on the WAIS-IV. Eight of the ten core subtests contributing to the calculation of FSIQ on the WAIS-IV have the same name as the core subtests contributing to FSIQ on the WAIS-III; the exceptions are Symbol Search and Visual Puzzles. In this third analysis, the WAIS-IV's Visual Puzzles subtest was treated as a missing or unmeasured variable for the WAIS-III sample (Keith & Reynolds, 2012), meaning its factor loading and unique variance were set to equal estimates obtained from the WAIS-IV sample. The Symbol Search subtest is included in the WAIS-III but is not a core subtest.

The test of invariance provides support or rejection of equivalence of scores across versions; however, it does not comprehensively address the question, of directionality or "To what extent was the publisher successful in

providing improvements in the areas of fluid reasoning, processing speed, and working memory?" The third analysis addresses this question by examining the average variance extracted (AVE) and construct reliability (CR) of each instrument. The AVE and CR statistics from each instrument may then be compared across versions. Thus, the AVE and CR of each instrument's first-order factor scores were examined in an effort to evaluate the level of support for the interpretation of each test's first-order factor scores as being reliable and valid (American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 1999) as well as to provide a comparison of support for the first-order factor scores (i.e., theoretical constructs of intelligence measured) across instruments. The AVE reflects the amount of common variance shared by subtests used to measure a theoretical construct while the CR reflects internal consistency (Anderson & Gerbing, 1988; Fleishman & Benson, 1987).

RESULTS

A key purpose of this study is to assist forensic and legal professionals in the identification of ID or Borderline intellectual functioning when an individual's score on the WAIS-III and WAIS-IV varies between the Borderline (i.e., FSIQ between 71 and about 79) and the ID range (i.e., FSIQ $\leq$ 70). This is analogous to the question "Which test provides the most valid FSIQ score when the criterion of reference is the test publisher's representation and measurement of intellectual functioning?"

The first analyses addresses this question by investigating how well each instrument's averaged covariance matrix fit each respective instrument's measurement and scoring model. In other words, does the WAIS-III's measurement and scoring model provide a better fit or does the WAIS-IV's measurement and scoring model: Which FSIQ score is more consistent with the test publisher's theoretical model to measure intelligence?

The first analysis tested three models. The first model, Model 1, tested the WAIS-III's measurement and scoring model as presented in Figure 1. As shown, the WAIS-III's measurement and scoring model is hierarchical in nature. The 11 subtests contributing to an individual's FSIQ are presented on the right side of Figure 1. These 11 subtests are then subsumed by one of two first-order factors: Verbal IQ (VIQ) or Performance IQ (PIQ). This is similar to an individual's receiving VIQ and PIQ scores on the WAIS-III. The VIQ and PIQ are subsumed by a general factor of intelligence or FSIQ.

Model 2, presented in Figure 2, is the measurement and scoring model of the WAIS-IV. The 10 subtests contributing to FSIQ on the WAIS-IV are also presented at the right side of its hierarchical model. Each of these 10 subtests is subsumed by (or load on) one of four first-order factors

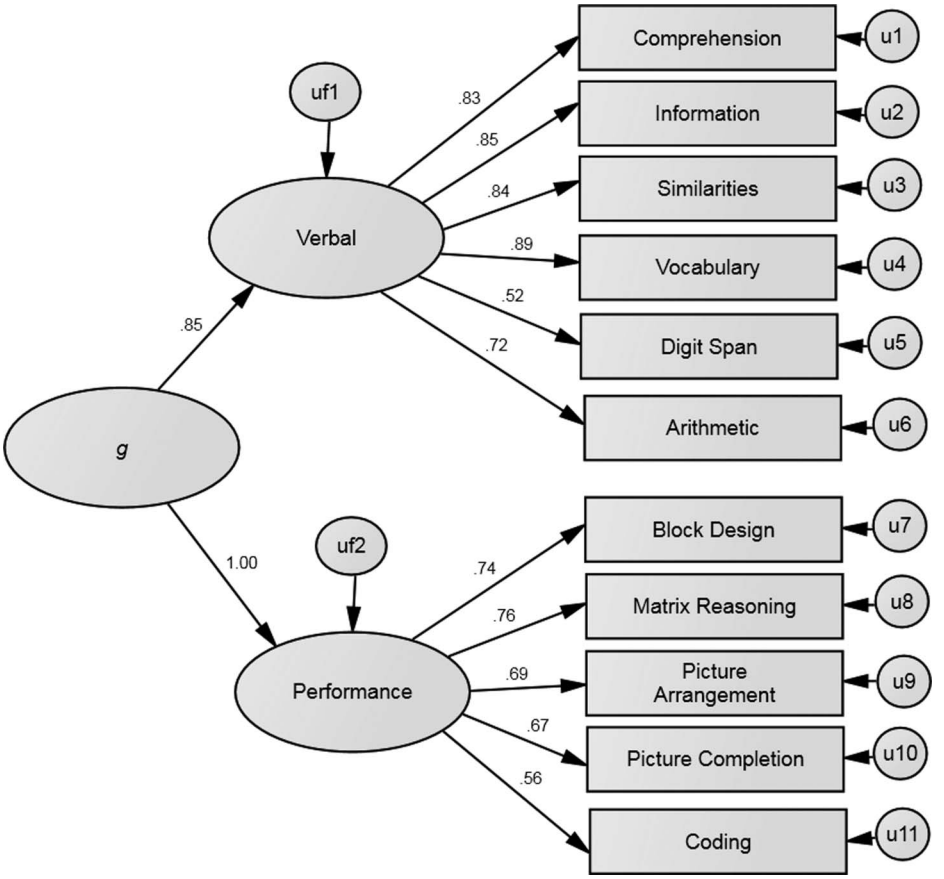

FIGURE 1 WAIS-III scoring measurement model.

(i.e., Verbal Comprehension, Perceptual Reasoning, Working Memory, or Processing Speed). When the WAIS-IV is administered, an individual obtains an index and scaled scores on each of these four first-order factors. An individual's FSIQ is derived from the sum of scaled scores associated with each of the four first-order factors.

To evaluate the fit of the data to each model, several indices of fit were examined; these include the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI). Values for these indices 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) was also included. The RMSEA is a fit index with values ranging from 0.00 to 1.00, with zero indicating a perfect fit. Values equal to or less than .05 indicate a good fit, and values up to .10 indicate a mediocre fit (Byrne, 2001). The Akaike Information Criteria (AIC) is another fit index that can be used to compare models. When two or more models are compared, the best model

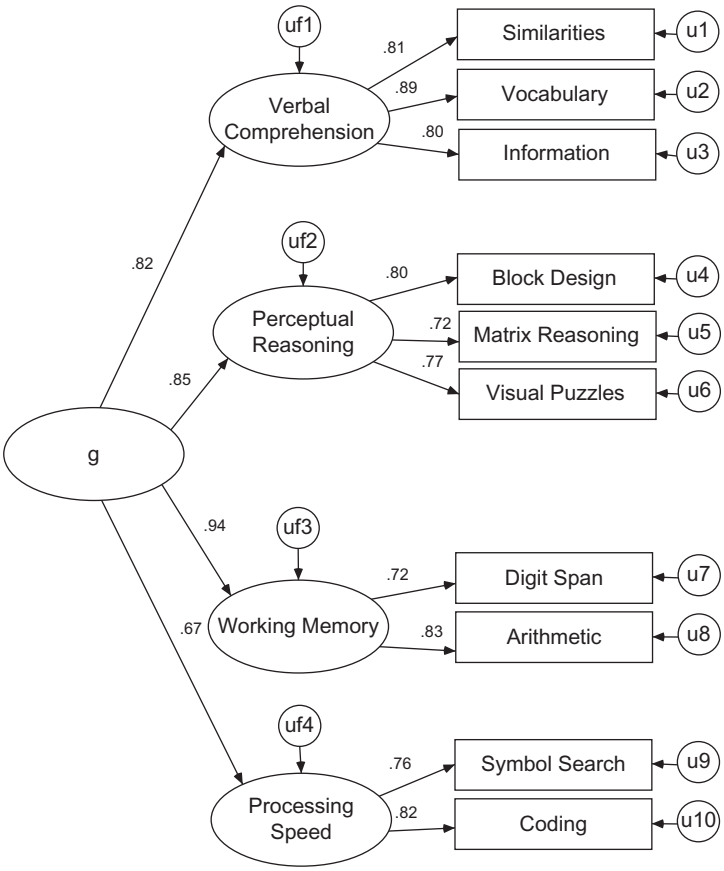

FIGURE 2 WAIS-IV scoring measurement model.

is the model with the lowest AIC. Finally, the chi-square was also calculated for each model. In general, when comparing models, the lower chi-square indicates a better fit. Because the models are not nested, the AIC provides the *best* index of change in fit across models (Keith et al., 2006).

A major change between the WAIS-III and the WAIS-IV was the incorporation of contemporary intellectual theory into the WAIS-IV's measurement and scoring models (Wechsler, 2008b). This can be seen in Figure 2 by the inclusion of four first-order factors, in contrast to the WAIS-III's VPIQ dichotomy.

The fit indices associated with each model are presented in Table 2. As shown in Table 2, the WAIS-IV provides better fit than the WAIS-III. The chi-square and RMSEA are both lower on the WAIS-IV; the CFI and TLI are both higher. Most important, however, both the AIC and chi-square on the WAIS-IV are also lower when compared to the WAIS-III. Taken together, the results indicate the WAIS-IV's measurement model provides a better fit to its normative sample when compared to the WAIS-III.

TABLE 2 Fit Statistics for Scoring Measurement Models

Model	χ^2 *(df)*	AIC	CFI	RMSEA	TLI
1. WAIS-III VIQ-PIQ	728.265 (43)	774.265	.955	.081	.943
2. WAIS-IV	261.216 (31)	309.216	.975	.064	.963
3. Four-factor WAIS-III	385.051 (41)	435.051	.977	.059	.970

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

Model 3 was evaluated in an effort to be comprehensive. Model 3 is the first-order *clinical* four-factor model of the WAIS-III converged into a measurement and scoring model. It is important to note that the clinical four-factor model is a *theoretical model* presented in the WAIS III's TM which does not contribute to the calculation of FSIQ (Wechsler, 2008b). In the present analysis, the clinical four-factor model was converted from a first-order model (a four-factor-only model) to a first- and second-order hierarchical model (four factors and FSIQ). Within the TM, the *clinical* four-factor model includes the 11 subtests contributing to FSIQ (via VIQ/PIQ) and two additional subtests, Symbol Search and Letter-Number Sequencing. Because this study is limited to the tests contributing to the calculation of FSIQ, the Symbol Search and Letter-Number Sequencing tests were not included in our analyses. Thus, Model 3 diverges from the *clinical* four-factor model in the TM in two ways: First is the inclusion of FSIQ and second is the exclusion of the two subtests from the clinical four-factor model that do not overlap with the traditional VIQ/PIQ scoring model. The *modified clinical* four-factor model of the WAIS-III is presented in Figure 3. It is also important to draw attention to the Processing Speed factor in Figure 3. This is because Processing Speed has only one indicator: the Coding subtest. Thus, the Coding subtest is isomorphic with the first-order Processing Speed factor, meaning the Processing Speed factor could be eliminated from Model 3. Tests of Model 3 with and without the Processing Speed factor were compared, and there was minimal to no change in model fit; most notably a change was not observed in either model's AIC. Thus, the decision was made to include the Processing Speed factor in Model 3 as presented in Figure 3 to assist the reader by providing a four-factor model that may be compared to the WAIS-IV's four-factor model (Model 2 in Figure 2).

The fit indices associated with the *modified clinical* four-factor WAIS-III are presented at the bottom of Table 2. As presented, the CFI and TLI are slightly better in the *modified clinical* four-factor WAIS-III than the WAIS-IV; however, the chi-square of the WAIS-IV is lower than the *modified clinical* four-factor WAIS-III and, most important, when comparing across models, the AIC was considerable lower on the WAIS-IV. Together these results indicate that the WAIS-IV's measurement model provides a better fit when compared to either the WAIS-III's measurement and scoring model or the WAIS-III's

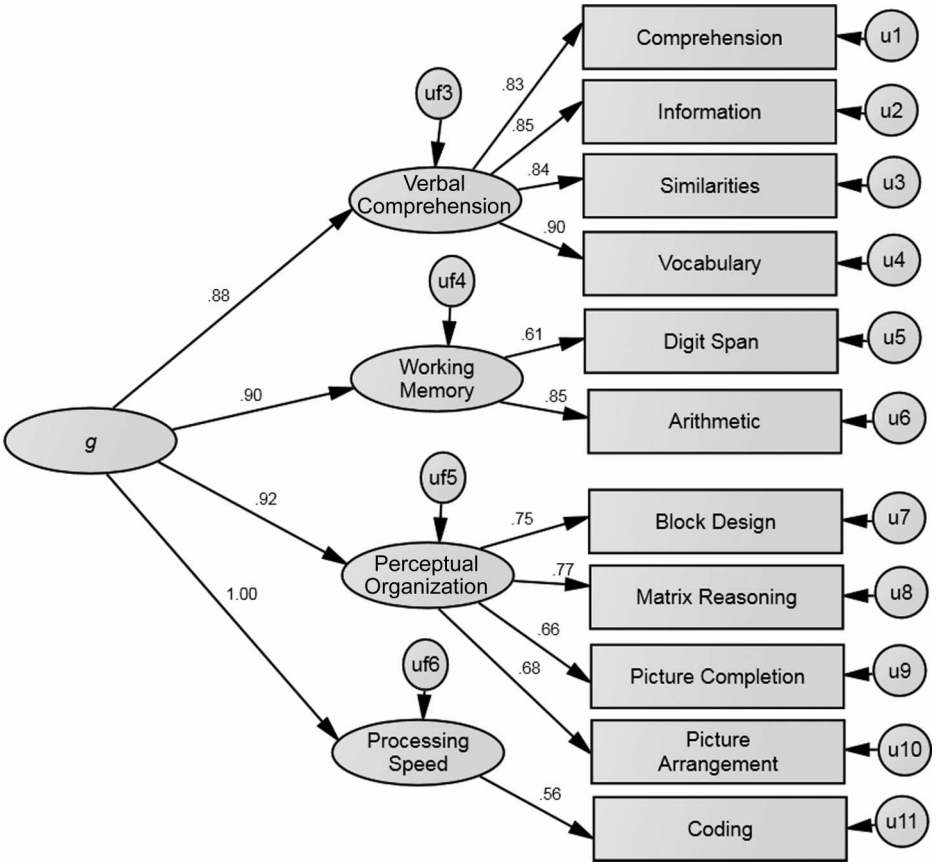

FIGURE 3 Theoretical four-factor WAIS-III measurement model.

modified clinical four-factor measurement and scoring model. Therefore, when the question arises as to which FSIQ score is more valid when the criterion is the test publisher's theoretical model to measure intelligence; the results indicate conclusively that the WAIS-IV's FSIQ is more valid than the WAIS-III's FSIQ score.

Contemporary research indicates the Cattell-Horn-Carroll (CHC) model of intelligence may provide the best *theoretical* measurement model to explain performance on the Wechsler scales (Benson et al., 2010; Keith et al., 2006). The second analysis examined both WAIS scales fit to a CHC measurement model. In this analysis, two additional models were developed and tested. Both of these measurement models were based on CHC theory. One model was developed for the WAIS-III; another model was developed and tested for the WAIS-IV. Each of these models included five first-order factors (see Benson et al., 2010 for additional information). Because we limited our analyses to the subtests contributing to FSIQ, it was not possible to generate results for either model (the models were under-identified). Thus, the

WAIS-III and WAIS-IV's measurement models and the *modified clinical* four-factor WAIS-III measurement and scoring model are the best and only models to compare the WAIS-III and WAIS-IV's measurement models when FSIQ is the outcome of interest. Thus, the results of this study indicate the WAIS-IV provides a better fit to its measurement and scoring model (theoretical structure) when compared to the WAIS-III.

The third analysis examined the invariance of the WAIS-III and WAIS-IV across their respective standardization samples. Results from the first analysis in the present study as well as previous research (Taub, 2001) suggest that a four-factor model provides the best fit to the WAIS-III's standardization data. The theoretically supported four-factor measurement and scoring model of the WAIS-IV was used to simultaneously test the invariance of both instruments. The standardization data from the WAIS-III served as the input data for the WAIS-III and the standardization data of the WAIS-IV served as the input data for the WAIS-IV.

The results presented in Table 3 suggest that scores derived from the WAIS scales are *not* equivalent across the third and fourth editions. The first test of invariance is a test of configural invariance. Within a test of configural invariance, all paths within the model are free or unconstrained. The results from the test of configural invariance indicates the unconstrained structural model had a statistically significant χ^2 value across both samples ($\chi^2 = 417.223$ [46], $p < .01$). The large sample sizes used in this study ($n = >2{,}200$) may be responsible for the significant χ^2; this is one of the main reasons why alternative fit statistics were developed (Bentler & Bonett, 1980; Marsh, Hua, Balla, & Grayson, 1998). An examination of the alternative fit indices suggests that the model provides a reasonable fit to the data (AIC = 541.223, CFI = .983, RMSEA = .042, TLI = .974). Thus, subtests do appear to have the *same structural configuration* and seem to measure *similar constructs* in both samples. The next test of invariance, metric invariance, is more restrictive. In metric invariance, the factor loadings of the first-order factors are fixed to be equal across instruments. The results from the test of metric invariance indicate large differences exist with respect to the magnitude of first-order factor loadings across the WAIS-III and WAIS-IV. According to the likelihood ratio test, model fit degraded when the first-order loadings

TABLE 3 Summary of Invariance Testing across the WAIS-II and WAIS-IV Standardization Samples

Model	Constraint	χ^2 *(df)*	$\Delta\ \chi^2$	Δdf	*p*	CFI	TLI	RMSEA	AIC
1.	Configuration	417.223 (46)	—	—	—	.983	.974	.042	541.223
2.	Measurement weights	484.882 (51)	67.659	5	<.001	.981	.973	.043	598.882

AIC = Akaike Information Criterion; CFI = Comparative Fit Index; RMSEA = root mean square error of approximation; TLI = Tucker Lewis Index.

were constrained to be equal ($\Delta\chi^2 = 67.659$ [5], $p < .01$). Thus, metric invariance is not tenable, and it is unnecessary to examine more stringent tests of invariance. A comparison of factor loadings across instruments found the greatest cross-sample differences on the Coding and Digit-Span subtests. Interestingly, the factor loading for the Coding subtest improved on the WAIS-IV when compared to the WAIS-III: .85 and .56, respectively. Similar improvements were found on the Digit-Span subtest: .73 and .61, respectively. Additionally, the standardized error variance accounted for by both tests was reduced on the WAIS-IV. These results mean that scores obtained on the WAIS-III are not equivalent to scores obtained on the WAIS-IV.

The fourth analysis examined the reliability and validity of each instrument's first-order factor scores via the AVE and CR statistics. The AVE (common variance) is calculated by obtaining the sum of squared factor loadings for a particular factor, then dividing this sum by itself plus the sum of standardized error variances. The formula for the CR (internal consistency) is similar, except factor loadings are summed before they are squared. Values greater than .5 for AVE and .7 for CR are considered minimally acceptable. AVE and CR estimates for WAIS-III and WAIS-IV constructs are presented in Table 4. An examination of Table 4 reveals that all values reach the minimum threshold for minimal acceptability with the exception of Processing Speed on the WAIS-III. As noted previously, the Processing Speed factor is measured by only one subtest, Coding. This low AVE and CR on Processing Speed is the result of construct under-representation on the WAIS-III when FSIQ is the outcome of interest; a minimum of two indicators is necessary for minimum construct representation. This finding is also supported by the WAIS-IV's TM (Wechsler, 2008b), which identifies Processing Speed as one of three cognitive areas targeted for "improvement" (p. 17) in the WAIS-IV, thus emphasizing the importance of Processing Speed as "dynamically related to mental capacity" (p. 18). The results from these analyses and the revision goals of the WAIS-IV indicate the WAIS-III does not provide an adequate measure of an individual's Processing Speed

TABLE 4 AVE and Construct Reliability for First-Order Factors

Construct		WAIS-III	WAIS-IV
Verbal Comprehension	AVE	.732	.709
	CR	.916	.880
Perceptual Reasoning	AVE	.515	.572
	CR	.809	.801
Working Memory	AVE	.549	.604
	CR	.703	.753
Processing Speed	AVE	NC	.654
	CR	NC	.790

AVE = average variance extracted; CR = composite reliability;
NC = not calculated.

when FSIQ is the outcome of interest. The AVE and CR statistics for the WAIS-IV's Processing Speed scores are acceptable: .654 and .790, respectively. Also worth noting is Verbal Comprehension's AVE for the WAIS-III and WAIS-IV: .732 and .709, respectively, as well as each instrument's CR: .916 and .880, respectively. The higher observed AVE and CR on the WAIS-III is a reflection of the WAIS-III's providing four subtests measuring Verbal Comprehension, compared to the WAIS-IV, which has three subtests measuring Verbal Comprehension. Results suggest that the AVE for Perceptual Reasoning is higher on the WAIS-IV when compared to the WAIS-III: .604 and .549, respectively, whereas the WAIS-III's CR is higher than the WAIS-IV's: .809 versus .801, respectively. The AVE and CR statistics for the scores on the WAIS-III and WAIS-IV's Working Memory factor indicate that the WAIS-IV's statistics are all higher.

DISCUSSION

The purpose of this study was to test the measurement and scoring models of the WAIS-III and the WAIS-IV in an effort to empirically answer the question "Which test provides a better measurement of intelligence, the WAIS-III or the WAIS-IV?" This question was addressed through four analyses. The first analysis tested the fit of the WAIS-III's measurement and scoring model against the WAIS-IV's measurement and scoring model. The results indicated that the WAIS-IV's model provided the best overall fit to the data. An alternative WAIS-III model was then tested: the WAIS-III's *modified clinical* four factor model (Model 3, Figure 3). The results from this analysis indicated the *modified clinical* four-factor model fit the WAIS-III's standardization data better than the WAIS-III measurement and scoring model based on the VPIQ dichotomy (Model 2, Figure 2). Yet, the WAIS-IV measurement and scoring model provided a better fit to its respective instrument's standardization data when compared to either of the WAIS-III's models. These results indicate that the WAIS-IV provides a better theoretical measurement of intelligence than the WAIS-III.

The second analysis investigated the fit of the WAIS-III and WAIS-IV to the CHC theoretical measurement model. The results of this analysis were unidentified, meaning the analyses could not be completed due to construct under-representation. The input data in the present study were limited to the core subtests from each instrument. The use of only core subtests in a five-factor CHC-based model resulted in the model's being unidentified.

The third analysis investigated the invariance of the WAIS-III and WAIS-IV simultaneously. The results from the first test of invariance, configural invariance, are supported and indicate the subtests on the WAIS-III

and WAIS-IV seem to have the *same structural configuration* and appear to measure *similar constructs* in both samples. Metric invariance is not supported, however. The results from the test of metric invariance indicate that scores derived from the WAIS-III are not equivalent to scores derived from the WAIS-IV. The finding of configural support does indicate that the comparison of scores between the WAIS-III and WAIS-IV is less similar to the comparison of *apples to oranges* and more similar to the comparison of *grapes to raisins*. Configural invariance indicates that the two instruments share similar properties (e.g., shared genetics or DNA); however, the lack of metric invariance indicates the intellectual constructs measured by the two instruments are on a different scale. The *grapes to raisins* analogy is used because raisins are more concentrated than grapes; so an ounce of raisins will have more antioxidants, sugar, and calories when compared to an ounce of grapes; therefore, an ounce of grapes is *not equivalent* to an ounce of raisins. Just as the FSIQ score on the WAIS-III is *not equivalent* to the identical FSIQ score WAIS-IV, they should not be directly compared because their concentration of tests and the intellectual constructs measured across test are different.

Given the revision goals for the WAIS-IV, the finding of a lack of equivalence across instruments is not unexpected for several reasons. The first reason is the publisher's use of a theoretical blueprint based on contemporary theories of intelligence and factor analytic research to develop the WAIS-IV: This was lacking in the development of the WAIS-III and resulted in the elimination of the VPIQ dichotomy in the WAIS-IV. The second reason is the extensive revisions necessary to meet the publisher's revision goals at the subtest level including administration; item activities, completion, modification, and presentation; scoring changes; and task modifications (see Table 1). The lack of metric invariance indicates the gain of three FSIQ points every 10 years (Flynn, 1984), does not necessary mean that the population is becoming smarter. In contrast, the observed differences in scores across instruments may be reflective of changes in administration, item and test stimuli modification, item scoring, as well as the departure from the traditional VPIQ dichotomy to a more research-based factor-analytic measurement model consistent with contemporary theories of intelligence (Beaujean & Osterlind, 2008; Brand, 1996; Kaufman, 2010; Rodgers, 1999; McGrew, 2010; Wechsler, 2008a, 2008b; Wicherts et al., 2004).

The fourth analysis investigated the AVE and CR of both instruments at the first-order factor level. This analysis provided an opportunity to investigate the extent to which the WAIS-IV publisher met its revision goals to improve the construct representation, reliability, and validity of the instrument's first-order factor scores in the areas of Fluid Reasoning, Processing Speed, and Working Memory. The other intellectual constructs measured

by the two instruments—crystallized intelligence and visual-spatial processing were also investigated. The results from these analyses are addressed at the first-order factor level: Perceptual Reasoning, Processing Speed, Verbal Comprehension, and Working Memory.

Perceptual Reasoning Index

As presented in Table 4, the Perceptual Reasoning Index (PRI) AVE is higher on the WAIS-IV when compared to the WAIS-III's AVE (magnitude of the difference: .57). In contrast, the CR is higher on the WAIS-III as compared to the WAIS-IV (magnitude of the difference: .08). This inconsistency may be due to the inclusion of four PRI subtests on the WAIS-III in contrast to only three subtests on the WAIS-IV; yet the values of the magnitude of difference between instruments clearly favors the WAIS-IV. The PRI is a combination of three subtests on the WAIS-IV: Block Design, Matrix Reasoning, and Visual Puzzles. Block Design and Visual Puzzles both measure visual-spatial processing ability; the Matrix Reasoning subtest is a measure of fluid reasoning. The factor loadings and standardized error variances favor the WAIS-IV over the WAIS-III, with the exception of a difference in magnitude in factor loading of .01 on the Visual Puzzles subtest, favoring the WAIS-III. Interestingly, the factor loading of the Matrix Reasoning subtest on the WAIS-III is .03 higher than the WAIS-IV; the standardized error variance is .06 higher on the WAIS-IV. Although the difference in magnitude of .03 is negligible, a key revision goal of the WAIS-IV is to provide a better measurement of fluid reasoning. The present results indicate the publisher was *not* successful in providing a more robust measure of fluid reasoning on the WAIS-IV. Possibly if the new subtest Figure Weights, a measure of fluid reasoning, was included as a core WAIS-IV FSIQ subtest, the instrument would provide a better measure of fluid reasoning. By extension, if Figure Weights was added as a core subtest, the PRI could possibly be divided into two factors: Visual Spatial Index and Fluid Reasoning Index. Taken together, the magnitude of differences between the WAIS-III and WAIS-IV statistics and loadings favor the WAIS-IV; however, the revision goal of improving the construct representation of fluid reasoning on the WAIS-IV is *not* supported by the present results, when the core subtests contributing to FSIQ score is the outcome of interest.

Processing Speed

AVE and CR estimates suggest scores on the WAIS-IV's Processing Speed factor and the associated subtests are more reliable and valid when compared to the WAIS-III (see Table 4). The primary reason for this is the WAIS-III has only one subtest contributing to the calculation of Processing Speed when

the intended outcome is FSIQ; thus, Processing Speed suffers from construct under-representation on the WAIS-III.

Verbal Comprehension

As presented in Table 4, the AVE statistic is higher on the WAIS-III than the WAIS-IV, by a magnitude of .02. The CR is also higher on the WAIS-III by a magnitude of .05 compared to the WAIS-IV. The values of the AVE and the CR increase with the inclusion of subtests within the equation. Therefore, it is assumed the primary reason for the observed difference in the AVE and the CR between the WAIS-III and WAIS IV is partially the result of four subtests measuring Verbal Comprehension (VCI) in the WAIS-III; in contrast, the WAIS-IV has only three subtests contributing to the VCI. Given the publisher's revision goals of improving the measurement of fluid reasoning, processing speed, and working memory, it is clearly evident that one subtest from the VCI was eliminated to meet this goal.

Working Memory

As presented in Table 4, the AVE and CR statistics for the Working Memory Index are both higher on the WAIS-IV. This result supports the publisher's goal of improving the measurement of working memory on the WAIS-IV.

SUMMARY

In summary, the theoretical measurement and scoring model of the WAIS-IV provides a better fit to the standardization data when compared to the fit of the WAIS-III's model to its standardization data. The results from tests of invariance indicate that the two instruments measure similar constructs; however, the same FSIQ score across instruments are not equivalent and should not be directly compared. Additionally, the observed differences in FSIQ between the WAIS-III and WAIS-IV are not due to the population's getting smarter by three FSIQ points each decade (e.g., Flynn, 1984); rather, the observed differences in FSIQ scores are most likely due to developmental, methodological, statistical, and theoretical improvements that result in the WAIS-IV's providing an overall better estimate of FSIQ (Beaujean & Osterlind, 2008; Brand, 1996; Kaufman, 2010; McGrew, 2010; Rodgers, 1999; Wicherts et al., 2004; Beaujean & Osterlind, 2004).

At the first-order factor level, the WAIS-IV provides a better measure of Processing Speed and Working Memory when compared to the WAIS-III. Perceptual Reasoning scores favor both instruments with the magnitude of difference favoring the WAIS-IV. Four of the 11 subtests on the WAIS-III provide measures of Verbal Comprehension, in contrast to three subtests

on the WAIS-IV. The additional subtest on the WAIS-III results in higher AVE and CR statistics but may also indicate the measurement of this construct is over-represented on the WAIS-III, when considering the Processing Speed factor is measured by just one subtest. The WAIS-IV provides a more balanced measure of intelligence at the first-order factor level, with specific improvements in the areas of processing speed and working memory. Thus, the results from the present study found the WAIS-IV to provide the most valid FSIQ score when the criterion of reference is the test publisher's representation and measurement of intellectual functioning.

Limitations

The present study was limited by the exclusion of supplemental tests from all analyses. This limitation, however, was consistent with the test publisher's analyses as well as necessary because most eligibility/ineligibility decisions are based on FSIQ, in contrast to subtest or factor scores, and practitioners generally administer the core subtests. Additionally, legal and psychological professionals generally make decisions based on an individual's FSIQ, in contrast to scores on core and supplemental subtests. This study is also limited to a comparison between the WAIS-III and the WAIS-IV's measurement and scoring model; thus, as new instruments are developed, practitioners must follow best practice and administer revised instruments as they become available. It is implicit within this study that only scores from valid test administrations be used in decision making and investigative activities. This study has much strength as well. This includes testing alternative measurement models for the WAIS-III and the WAIS-IV as well as the application of confirmatory factor analysis and invariance testing via structural equation modeling.

Implications for Practitioners

This study found the WAIS-IV's FSIQ score is more valid than the WAIS-III's FSIQ score when the outcome criterion is the test publisher's theoretical model to measure intelligence. These results provide practitioners with empirical support for the WAIS-IV as a *technological improvement* over the WAIS-III. This finding along with the publisher's revision goal of improving the measurement of intelligence at the low end of the instrument (e.g., FSIQ $\leq$ 69; Wechsler, 2008b) provides additional support for the WAIS-IV's FSIQ score as being more valid and reliable score when compared to the WAIS-III's FSIQ score to discriminate between the identification of Borderline and ID. If an inmate were administered both versions of the Wechsler scales and scores differ, forensic psychologists and legal professionals should assign more weight to the FSIQ obtained from the WAIS-IV and consider it to provide a score that is more valid, reliable, and consistent with the publisher's

theoretical model to measure intelligence when compared to the WAIS-III's FSIQ. Consideration of empirical evidence will help protect against injustices when addressing the intelligence criterion in Atkins cases.

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.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411–423.

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

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

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22*(1), 121–130. doi:10.1037/a0017767

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. doi:10.1037/0033-2909.88.3.588

Beaujean, A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the national longitudinal study of youth 79 children and young adults' data. *Intelligence*, *36*, 455–463.

Brand, C. R. (1996). *The g factor: General intelligence and its implications*. Chichester, UK: Wiley.

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, NY: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 122–130). Needham Heights, MA: Allyn & Bacon.

Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement*, *47*, 925–939.

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

Flynn, J. R. (2007). *What is intelligence?* 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. (2010). "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn effect. *Journal of Psychoeducational Assessment*, *28*, 382–398.

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.

Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 758–799). New York, NY: Guilford.

Marsh, H. W., Hua, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–220.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1–10. doi:10.1016/j.intell.2008.08.004

McGrew, K. S. (2010). The Flynn effect and its critics: Rusty linchpins and "lookin' for g and Gf in some of the wrong places." *Journal of Psychoeducational Assessment*, *28*, 448–468.

Rodgers, J. L. (1999). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*, 337–356.

Rodgers, J. L., & Wanstrom, L. (2007). Identification of a Flynn effect in the NLSY: Moving from the center to the boundaries. *Intelligence*, *35*, 187–196.

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. (1939). *Wechsler-Bellevue Intelligence Scale*. New York, NY: The Psychological Corporation.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. New York, NY: The Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997a). *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.

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

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale-Fourth Edition*. San Antonio, TX: Pearson Assessment.

Wechsler, D. (2008b). *Wechsler Adult Intelligence Scale-Fourth Edition*: Technical and interpretive manual. San Antonio, TX: Pearson Assessment.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P, van Baal, G. C. M., Boomsma, D. I., . . . Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509–537.