Is the Cattell–Horn–Carroll-Based Factor Structure of the Wechsler Intelligence Scale for Children—Fifth Edition (WISC-V) Construct Invariant for a Representative Sample of African–American, Hispanic, and Caucasian Male and Female Students Ages 6 to 16 Years?

Caroline Scheiber¹

Received: 13 May 2016 / Accepted: 24 August 2016 / Published online: 8 September 2016
© American Academy of Pediatric Neuropsychology 2016

Abstract Since the inception of testing, the examination of test bias has been discussed and debated in the literature. As new tests are developed or old tests are revised, part of the standardization process involves the examination of gender and ethnic item bias. The exploration of construct invariance across different ethnic and gender groups, however, is not part of the test construction or validation process. This study investigated factorial invariance of the Cattell–Horn–Carroll (CHC)-based factor structure of the most recent version of the Wechsler tests—the Wechsler Intelligence Scale For Children—Fifth Edition (WISC-V; Wechsler 2014a)—across a representative sample of African–American, Hispanic, and Caucasian male and female students ages 6–16 years ($N=2637$). Specifically, confirmatory factor analysis was used to explore whether increasing sets of equality constraints fit the underlying theoretical structure of the WISC-V equally well. Results showed that all five CHC broad abilities, in addition to the general factor g , demonstrated structural invariance for all six ethnic/gender groups. Findings from this study suggest that, despite mean score differences, neuropsychologists can be confident that the WISC-V factor structure is equivalent for the different subgroups and scores are, thus, equally meaningful for African–American, Hispanic, and Caucasian male and female students. These results are of value, given the frequent use of intelligence tests, especially of the WISC-V, when making important and, oftentimes, life-changing decisions regarding diagnosis and access or denial to special programs and classes.

Keywords Wechsler Intelligence Test for Children—Fifth Edition · Construct invariance · Ethnicity · Gender · Bias

Intelligence tests are frequently used in determining eligibility and access to special programs and classes (e.g., Weiss et al. 2006, 2015). Despite much discussion and debate in the literature with regard to the fairness of test instruments, those studies that have looked at test bias across different population subgroups have generally found no gender or ethnic group bias (e.g., Edwards & Oakland, 2006; Keith 1999; Keith et al. 1999; Scheiber 2016a, b; Scheiber & Kaufman, 2015; Reynolds et al. 2015; Trundt 2013).

According to Reynolds and Lowe (2009), a test is biased “if a test design, or the way results are interpreted ..., systematically disadvantages certain groups ... over others [it is] a systematic error in the [design of the test that results in the erroneous] measurement of a psychological attribute as a function of membership in one or another ... subgroup” (p. 333). Similarly, the American Educational Research Association (AERA) defines that a test is biased “when deficiencies in the test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups” (AERA, American Psychological Association, and National Council on Measurement in Education 1999, p. 74).

Some researchers and clinicians still interpret mean score differences between subgroups as evidence of test bias, echoing arguments that were raised in the distant past (e.g., Williams 1971). This definition of bias rests on the premise that statistically significant differences in mean scores indicate that the test is biased against the subgroup that scores lower. Historically, differences in IQs, specifically across different ethnic groups, have persisted for decades and continue to exist

✉ Caroline Scheiber
caroline.scheiber@dhha.org

¹ Denver Health Medical Center, Denver Colorado, USA

today (Kaufman et al. 2016; Weiss et al. 2015). For example, on the most recent version of the Wechsler tests—the Wechsler Intelligence Scale for Children—Fifth Edition (WISC-V; Wechsler, 2014a)—Caucasians averaged a standard score of 103.5, Hispanics averaged 94.4, and African-Americans averaged 91.9 on the Full Scale IQ (FSIQ). Asians outperformed all three groups averaging 108.6 (Weiss et al. 2015).

Mean gender differences have also been found to vary across the lifespan but generally tend to be trivial compared to ethnic group differences and often are close to zero on global scales (e.g., Kaufman and Kaufman 2004; Kaufman et al. 2016; Keith et al. 2008; Reynolds et al. 2008). At the school-age level, the most consistent findings on specific abilities are as follows: (1) males outperform females in visual processing (Kaufman and Kaufman 2004; Keith et al. 2011; Jirout and Newcombe 2015; Palejwala and Fine 2014); (2) very young females outperform very young males in crystallized ability (Kaufman and Kaufman 2004; Keith et al. 2011); and (3) females outperform males in processing speed (Keith et al. 2011). On the WISC-V, researchers have observed no meaningful gender differences on the FSIQ and only trivial gender differences on most indexes, except on the Processing Speed Index. Generally speaking, gender differences on cognitive tests tend to be small, with the exception of processing speed, which shows a decided female advantage across the lifespan (Kaufman et al. 2016; Keith et al. 2011; Lichtenberger and Kaufman 2013).

Despite the frequent emphasis in research on mean ethnic differences on global scores and mean gender differences on specific abilities, such discrepancies may occur for an array of reasons, most of which relate to home environment (Weiss et al. 2015), and they tell us little about test bias or test fairness. Indeed, this approach to test bias is flawed; the simple comparison of mean scores is not a statistically sound way of determining a test's psychometric properties, as a number of researchers have highlighted (e.g., Reynolds and Lowe 2009; Reynolds and Keith 2013). Instead, bias is better defined by the absence or presence of a systematic error in the design or *construct* of the test instrument.

Test developers devote a significant amount of time, ensuring adequate test validity and reliability to avoid systematic errors and flaws in the design of the test instrument. However, exploring construct consistency of the test scores across different gender or ethnic groups—to ensure fairness in the constructs that they measure across diverse populations—is not part of the test development or validation process.

A few researchers have investigated the factorial invariance of contemporary intelligence tests for males versus females (e.g., Keith et al. 2008, 2011; Reynolds et al. 2008, 2015), and a handful have done so across different ethnic groups (Edwards & Oakland, 2006; Keith 1999; Keith et al. 1999; Scheiber 2016a; Trundt 2013). However, no study has

investigated factorial invariance across different ethnic/gender groups together. Furthermore, no study has yet explored factorial invariance by gender or ethnicity for the WISC-V—the most recent version of the Wechsler tests for children. The primary purpose of the present study was, therefore, to explore test score bias, in the form of differential construct validity, of the WISC-V, using confirmatory factor analysis (CFA). The goal was to determine whether there was bias in the assessment of theoretical constructs for Caucasian, African-American, and Hispanic male and female students ages 6–16.

The Wechsler Intelligence Scale for Children—Fifth Edition

The Wechsler tests are the most frequently used assessment instruments for children when it comes to diagnosis and the determination of access to special programs or classes (Kaufman et al. 2016). One theoretical model that is consistent with the conceptual foundation of the WISC-V is the popular Cattell–Horn–Carroll (CHC) theory of intelligence (Schneider and McGrew 2012), a comprehensive and empirically validated approach to the measurement of intelligence. It hypothesizes that the structure of intelligence rests on a hierarchical model that is composed of three strata. Stratum I consists of over 40 narrow abilities, which measure a person's "level of mastery, along a difficulty scale..., speed with which the individual performs tasks..., or the rate of learning in learning and memory tasks" (Carroll 1993, p. 124). Stratum II measures several broad abilities, and Stratum III represents the general intelligence factor (*g*), a global construct which is hypothesized to be at the core of all broad and narrow abilities.

Similarly to most contemporary tests of intelligence, the WISC-V scales correspond to CHC broad abilities. Indeed, each of the five WISC-V primary index scores corresponds to a CHC ability: the Verbal Comprehension Index corresponds to crystallized ability (*Gc*), the Visual Spatial Index corresponds to visual processing (*Gv*), the Fluid Reasoning Index corresponds to fluid problem solving (*Gf*), the Processing Speed Index corresponds to processing speed (*Gs*), and the Working Memory Index corresponds to short-term memory (*Gsm*). In addition to those five broad abilities, the WISC-V also offers the option to yield a sixth (supplementary) scale, Symbol Translation, which is representative of long-term storage and retrieval (*Glr*). The Full Scale IQ (FSIQ) corresponds to the Stratum III general ability factor *g*.

The WISC-V publisher provided evidence for reliability and validity of the test instrument (Wechsler 2014b); however, no studies exist that provide evidence for the construct invariance of the WISC-V for different ethnic or gender groups.

Present Study

When the WISC-V was developed, the publisher paid extra attention to minimizing bias against certain subgroups (e.g., special populations, second language learners) by using expert opinions and applying statistical analysis. Furthermore, CFA was employed to evaluate the quality of the model and verify the overall factor structure of the WISC-V (Wechsler 2014b). The CHC theory, which conforms closely to the factor structure of the WISC-V, has been found to be applicable to diverse ethnic groups (Flanagan et al. 2014; Scheiber 2016a, b; Schneider and McGrew 2012). This study seeks to extend this line of research by determining whether the applicability of CHC across ethnicities is equally valid for males and females within each ethnic group. At the root of the research is to determine whether the five CHC constructs—*Gc*, *Gf*, *Gv*, *Gsm*, and *Gs*—that underlie the WISC-V are invariant for the ethnic/gender samples (e.g., Hispanic males, Caucasian females). If the constructs are *not* invariant, then the goal is to identify whether bias exists against any particular subgroup. Of note is that *Glr* was not included in the analysis. This is because the three subtests that comprise *Glr* are experimentally dependent. In other words, they all are based on the same Symbol Translation test—one is immediate learning, one is delayed recall, and one is recognition of symbols. For all other five scales, the component subtests are independent subtests in terms of their administration. A construct should only be measured if one has at least two independent measures of the construct.

Methods

Participants

The sample consisted of children and adolescent ages 6 through 16 from the norming sample of the WISC-V. It included a total of $N=2637$ students with $n=777$ Caucasian males, $n=830$ Caucasian females, $n=188$ African–American males, $n=221$ African–American females, $n=308$ Hispanic males, and $n=313$ Hispanic females. The sample was stratified on important background variable (e.g., ethnicity, gender, parent educational attainment, age, and geographic region) according to 2012 US Census data. In other words, the normative sample has a close correspondence with the 2012 US Census proportions in terms of age group, gender, ethnicity, geographic region, and parent educational level. The latter served as a proxy for socioeconomic status. Each age group had appropriate numbers of the different levels of parental educational attainment, as they correspond to 2012 US Census. Parental educational attainment (as proxy for socioeconomic status) was, thus, stratified across and within groups. More details on the sample and the psychometric

properties of the WISC-V can be found in the WISC-V Technical and Interpretative Manual (Wechsler 2014b).

Statistical Procedure

Factorial Invariance Using Multi-Group Mean and Covariance Structure In order to explore whether the factor structure of the WISC-V is the same for Caucasian, Hispanic, and African–American male and female students, factorial invariance was assessed. Specifically, the variable structure of the WISC-V that included all 16 core and supplementary subtests (Wechsler 2014b) was explored to determine whether it is invariant across the six different ethnic/gender groups. The 16 subtests used comprised the variables entered into the CFA: (a) the WISC-V subtests are organized according to five core CHC ability measures: *Gc*, *Gf*, *Gsm*, *Gv*, and *Gs*, and (b) the factorial invariance of the first-order factors was assessed in each analysis. The first-order variables refer to the five CHC latent factors (*Gc*, *Gf*, *Gsm*, *Gv*, and *Gs*). The invariance of the second-order *g*-factor (FSIQ), which is hypothesized to underlie all of the subtests that compose the WISC-V, was also explored.

Analytical Steps In order to control for the effects of age, age standardized scores were used to conduct this analysis. To explore factorial invariance of test scores across groups, multi-group confirmatory factor analysis (MG-CFA) based on a mean and covariance structure (MG-MACS) was applied (Reynolds and Keith 2013). The equivalence of the WISC-V factor structure was assessed for Caucasian, African–American, and Hispanic male and female students ages 6 through 16. This method was used to explore whether the same construct was being measured across different groups. Amos software version 20 (Arbuckle 1995–2011) was used for all analyses.

Testing for measurement invariance using MG-MACS required the setting of increasingly restrictive sets of equality constraints: (1) configural invariance, (2) metric invariance, (3) intercept invariance, and (4) residual invariance (Meredith 1993).

To establish *configural* invariance, the same factor structure of free and fixed loadings was applied to all subgroups. To balance the factors and scale them, one reference indicator for the second-order factor and each first-order factor was fixed to 1. The other variances and covariances were allowed to vary freely. Whereas the intercepts could vary freely, the latent factor means were fixed to 0. If configural invariance was established, *metric* invariance could be assessed. Here, in order to have all corresponding factor loadings equal across the six groups, constraints were added to the previously free first-order factor loadings. This ensured that a one unit increase in a specific factor for one group would result in the same unit increase for the other group. If metric invariance

was established, it could be concluded that the relation between subtests and factors was the same for the different subgroups (Reynolds and Keith 2013).

The next step included the establishment of *intercept invariance*. To test for intercept invariance, all corresponding subtest intercepts (means) were constrained to be equal across the groups. In other words, this specification explored whether the same subtest score across the different subgroups resulted in the same latent CHC factor ability score. In other words, it was assessed whether a specific subtest score would mean the same for one subgroup as it would for another subgroup. When all constraints were added and the fit index did not degrade, it was concluded that the test scores measured the same underlying construct across all subgroups (Reynolds and Keith 2013). Alternatively, if intercept invariance was *not* established, then this was an example of bias, because the same subtest scores would mean something different for members of different subgroups. In other words, the subtest did not measure the ability that it was supposed to measure. In such a case, observed scores on the tests would not be based on ability but would be the result of psychometric test score bias due to lack of construct validity (Reynolds and Keith 2013).

Partial strong factorial invariance means that measurement invariance for a specific factor could not be successfully established simply because one particular subtest was not invariant across the groups. In other words, there is something about this particular subtest that cannot solely be explained by the ability factor. This subtest would have to be discovered, and the constraints of that particular subtest intercept would then be freed and partial factorial invariance would be established (Reynolds and Keith 2013).

There was an option to also investigate *residual invariance*, which refers to the equality constraints of the error variances. However, as several researchers have suggested (e.g., Byrne et al. 1989; Reynolds and Keith 2013), residual invariance is *not* a necessary prerequisite in order to establish construct invariance. Thus, this study did not assess residual invariance.

Fit Indexes The likelihood ratio test (χ^2), the root-mean-square error of approximation (RMSEA), and comparative fit index (CFI) were used to assess model fit. In order to conclude very good model fit, results should show a non-significant $\Delta\chi^2$, a CFI value of at least 0.95 (Cheung and Rensvold 2002), and a RMSEA value of less than 0.05 or less (Reynolds and Keith 2013). Adequate to good model fit would be reflected in a CFI value of at least 0.90 and a RMSEA value of 0.08 or less (Bentler 1990; Crocker et al. 2003; Van der Vegt et al. 2001). Furthermore, comparison of goodness-of-fit indexes should be non-significant for $\Delta\chi^2$, <0.01 for ΔCFI , and <0.02 for ΔRMSEA (Cheung and Rensvold 2002). As $\Delta\chi^2$ can easily result in a significant degradation in model fit as it detects minor or inconsequential differences, more weight was placed on the CFI and RMSEA

values to conclude model fit for the invariance analysis (Byrne, 1989; Chen et al. 2005; Reynolds and Keith 2013). Furthermore, to evaluate the degradation in model fit, ΔCFI and the ΔRMSEA were used. RMSEA had to be multiplied by its square root of 6 due to multiple groups (Steiger 1998).

Results

Descriptives

Means and standard deviations for the WISC-V indexes and subtests for the different ethnic and gender groups are reported in Table 1. Assumptions of normality, interval or ratio level of measurement, independence, and linearity underlie CFA. For the total sample ($N=2637$), skewness ranged from -0.108 (block design) to $+0.115$ (similarities) and was, therefore, far from the ± 2.0 cutoff. Kurtosis ranged from -0.174 (visual puzzles) to 0.428 (symbol search) and was, therefore, far from the ± 7 cutoff (Meyers et al. 2013). Skewness and kurtosis were also evaluated for the ethnic and gender groups separately and were normal. Based on skewness and kurtosis data, data points for the total sample and for each ethnic group were normally distributed. Furthermore, all data are derived from raw scores on subtests that were normalized and then standardized to have a mean = 100 and $SD = 15$ and, therefore, fall at an interval level of measurement (Wechsler 2014b). Independence was met due to the stratified random sampling procedures, and linearity was met as well for the total sample (Wechsler 2014b).

Frequency distributions were examined to uncover missing data. Missing data was small (less than 5 % of total data points). Out of $N=2637$, the Similarities, Symbol Search, Digit Span, Picture Span, and Coding had between 8 and 13 missing data points and Vocabulary, Block Design, Visual Puzzles, Matrix Reasoning, Figure Weights, and Coding all had between 2 and 5 missing data points. All missing data points were handled using the mean scaled score as best estimate of the missing scaled score, relying on its scale membership. In other words, if a Verbal Comprehension Index (VCI) subtest was missing, the means of the other VCI subtests (including supplemental) were used to estimate the missing score. In that sense, the missing data points for, for example, the Vocabulary and Similarities subtests were replaced by calculating the mean score of the Comprehension and Information subtests.

Creation of the Model

Before factorial invariance could be established, the model that best fit the data for all samples, including the total sample ($N=2637$), had to be identified. First, a model was hypothesized to fit the data (“original model”) and then that model was

Table 1 Means and standard deviations (SD) for each CHC-factor and WISC-V subtest for Caucasian, African–American, and Hispanic females

WISC-V indexes and subtests	Caucasian females		African–American females		Hispanic females	
	Mean	SD	Mean	SD	Mean	SD
	(n = 830)		(n = 221)		(n = 313)	
Verbal Comprehension Index/Gc	99.95	14.78	91.60	12.42	93.87	13.41
Comprehension	10.89	2.87	9.05	2.54	9.03	2.98
Information	10.18	2.58	8.47	2.57	8.59	2.55
Similarities	10.78	2.84	8.81	2.55	9.02	2.66
Vocabulary	10.69	2.72	8.08	2.56	8.70	2.72
Visual Spatial Index/Gv	102.97	13.85	90.95	13.67	95.83	12.75
Block design	10.56	2.79	8.38	2.95	9.29	2.52
Visual puzzles	10.54	2.79	8.40	2.69	9.21	2.75
Fluid Reasoning index/Gf	103.2	13.66	93.54	14.12	96.34	13.43
Arithmetic	10.47	2.78	8.68	2.64	8.92	2.47
Figure weights	10.45	2.73	8.68	2.75	9.21	2.69
Matrix reasoning	10.63	2.85	9.11	2.98	9.55	2.70
Picture concepts	10.31	2.92	9.20	3.21	9.39	2.89
Working Memory Index/Gsm	103.56	14.05	97.35	14.22	95.00	13.90
Digit span	10.56	2.82	9.53	2.79	9.00	2.70
Letter–number sequencing	10.69	2.85	9.15	2.99	9.20	2.69
Picture span	10.65	2.88	9.57	2.92	9.32	2.89
Processing Speed Index/Gs	104.82	14.67	99.67	15.36	101.70	13.70
Cancellation	10.49	3.09	10.35	3.15	10.16	2.99
Coding	11.02	2.86	10.23	3.04	10.68	2.79
Symbol search	10.69	2.97	9.53	3.26	9.92	2.89
Full Scale IQ/g	104.97	13.55	92.93	13.22	95.28	12.45
	(n = 777)		(n = 188)		(n = 308)	
Verbal Comprehension Index/Gc	104.27	14.67	92.57	15.05	95.92	13.95
Comprehension	10.35	2.99	8.71	2.74	9.10	2.99
Information	11.12	2.87	9.03	2.98	9.45	3.05
Similarities	10.71	3.03	8.80	3.04	9.31	2.92
Vocabulary	104.05	15.12	90.39	13.76	98.54	14.09
Visual Spatial Index/Gv	106.60	2.97	8.49	2.89	9.82	2.83
Block design	10.85	3.08	8.11	2.78	9.65	2.89
Visual puzzles	10.85	2.89	8.41	3.00	9.16	2.76
Fluid Reasoning index/Gf	103.37	15.09	93.08	14.89	96.61	14.41
Arithmetic	10.71	3.03	8.76	2.64	9.46	2.99
Figure weights	10.84	3.03	8.74	2.79	9.46	2.84
Matrix reasoning	10.32	3.01	8.90	3.16	9.41	2.93
Picture concepts	10.51	3.02	9.15	2.82	10.02	3.06
Working Memory Index/Gsm	101.92	15.22	94.09	13.59	95.77	15.26
Digit span	10.55	3.06	9.06	2.75	9.21	2.99
Letter–number sequencing	10.52	3.02	8.76	2.85	9.06	3.05
Picture span	10.12	3.02	8.88	2.66	9.35	2.93
Processing Speed Index/Gs	98.31	14.48	94.39	15.69	97.36	13.93
Cancellation	9.63	2.98	9.96	3.33	10.35	2.95
Coding	9.37	2.92	9.03	2.88	9.28	2.92
Symbol search	10.01	2.82	8.95	3.29	9.76	2.68
Full Scale IQ/g	103.33	14.86	91.22	13.63	95.50	13.95

Data and table copyright Pearson

modified based on theory and data to try to improve it. For this study, the original model was based on the CHC-based factor structure as it is demonstrated in Model 5a in the WISC-V Technical and Interpretative Manual (Wechsler 2014b). Model 5a is a five-factor model that corresponds to the five indexes of the WISC-V: (1) the VCI (equivalent to *Gc*), the Fluid Reasoning Index (equivalent to *Gf*), the Working Memory Index (equivalent to *Gsm*), the Processing Speed Index (equivalent to *Gs*), and the Visual Processing Index (equivalent to *Gv*). This five-factor model showed good model fit for the total sample ($\chi^2(96)=5.403$, $p<.05$; CFI=0.97; RMSEA=0.05).

Table 2 shows the model fit for the present sample. It shows the original CHC classification of all WISC-V subtests based on Model 5a of the WISC-V Technical and Interpretative Manual (Wechsler 2014b) for the total sample and for each ethnic/gender group separately. It also demonstrates the data- and theory-driven model. Based on modification indexes and theory, the WISC-V Arithmetic subtest was cross-loaded with *Gc* and *Gsm* in addition to the original loading on *Gf*. With the changes to the model, fit indexes improved for the total sample: the values of CFI improved from 0.98 to 0.99, and RMSEA values improved from 0.034 to 0.030 for the total sample. Similarly, the model fit for each individual subgroup also improved with the modification of the model: the new values of CFI for the subgroups ranged from 0.98 to 0.99, and the values of RMSEA were all equal to or less than 0.04. Arithmetic heavily loads on *Gc* and *Gsm*, as it requires the ability to use prior knowledge base in the application of novel math problems, read the question, and solve it both verbally and quantitatively, especially in higher grades (Wendling and Mather

2009). Arithmetic also requires the ability to keep information in the short-term memory and manipulate this information while solving the math task (Keith et al. 2006; Weiss et al. 2013). In fact, according to studies conducted with the WISC-V, Arithmetic tends to load on *Gsm* in a four-factor structure but on *Gf* in a five-factor structure (Wechsler 2014b). It was, therefore, decided to maintain these cross-loadings as it made sense based on theory and improvement of fit indexes.

This final model, in fact, corresponded to the final CHC factor model as demonstrated in the WISC-V Technical and Interpretative Manual (Wechsler 2014b, Fig. 5.1). Similarly to the current analysis, the WISC-V authors also found that allowing Arithmetic to load on Working Memory (*Gsm*) and Crystallized Ability (*Gc*) in addition to Fluid Reasoning (*Gf*) significantly improved fit for the total sample. Figure 1 shows the final five-factor CHC structure of the WISC-V as it was used for the subsequent analyses.

Invariance Analysis

Table 3 demonstrates the model fit for the multi-group comparison. The configural invariance analysis yielded a significant chi-square value ($\chi^2(582)=1.459$, $p<.001$); however, this was expected given the number of comparisons that were conducted for the six subgroups. More value was placed on the CFI and RMSEA values, which showed good model fit for configural invariance (CFI=0.98 and RMSEA=0.032). Based on the CFI and RMSEA values, configural invariance was established. As more constraints were added to the model for the metric invariance analysis, as expected, the chi-square

Table 2 Model fit indexes for confirmatory factor analysis (CFA) original and data- and theory-driven models for the total sample ($N=2637$) and the six ethnic and gender groups (Caucasian females $n=830$, Caucasian males $n=777$, African-American females $n=221$, African-American males $n=188$, Hispanic females $n=313$, and Hispanic males $n=308$)

Form and model	χ^2	<i>df</i>	<i>P</i> value	CFI	RMSEA	RMSEA 90 % CI
Original model						
Total sample	419.949	97	<.001	0.98	0.034	0.031–0.037
Caucasians M	182.935	97	<.001	0.98	0.034	0.026–0.041
Black M	136.114	97	.005	0.97	0.046	0.026–0.064
Hispanics M	120.244	97	.055	0.99	0.028	0.000–0.043
Caucasian F	142.014	97	<.001	0.99	0.024	0.015–0.032
Black F	135.484	97	.006	0.97	0.042	0.023–0.059
Hispanic F	132.061	97	.010	0.98	0.034	0.017–0.048
Data- and theory-driven model						
Total sample	323.662	95	<.001	0.99	0.030	0.027–0.034
Caucasians M	145.800	95	.001	0.99	0.026	0.017–0.034
Black M	123.922	95	.025	0.98	0.040	0.015–0.059
Hispanics M	112.395	95	.107	0.99	0.024	0.000–0.041
Caucasian F	125.579	95	.019	0.99	0.020	0.008–0.028
Black F	128.623	95	.012	0.98	0.040	0.020–0.057
Hispanic F	127.487	95	.015	0.98	0.033	0.015–0.047

Data and table copyright Pearson

CF comparative fit index, RMSEA root-mean-square error of approximation

Fig. 1 Proposed five-factor CHC structure of the Wechsler Intelligence Scale for Children—Fifth Edition (WISC-V)

value remained significant ($\chi^2 (647) = 1.435, p < .001$). However, no changes in CFI and RMSEA values occurred. Thus, metric invariance for the first-order factors was established. Similarly, no changes in CFI and RMSEA were observed when the second factor loadings were added, although a significant chi-square value was detected ($\chi^2 (667) = 1.422, p < .001$). However, based on the good CFI and RMSEA values, metric invariance was established.

For the intercept invariance analysis, the subtest intercepts were constrained to be equal across the groups. Although the chi-square value remained significant ($\chi^2 (722) = 1.886,$

$p < .001$), CFI and RMSEA values provided evidence for very good model fit (CFI = 0.96 and RMSEA = 0.04). Furthermore, the changes observed in CFI values were small ($\Delta\text{CFI} = 0.02$) and in RMSEA values non-significant ($\Delta\text{RMSEA} < 0.02$). Finally, the second-order intercept constraints were added, which again resulted in a significant chi-square value ($\chi^2 (747) = 2.201, p < .001$). However, once again, the CFI and RMSEA values showed very good model fit (CFI = 0.95 and RMSEA = 0.05), and the changes observed in CFI and RMSEA values were trivial or non-significant ($\Delta\text{CFI} = 0.01; \Delta\text{RMSEA} < 0.02$). Based on the CFI and RMSEA values,

Table 3 Model fit indexes and nested comparison confirmatory factor analysis (CFA) models: Caucasian males, Caucasian females, Hispanic males, Hispanic females, African-American males, and African-American females

Form and model		χ^2	df	$\Delta\chi^2$	Δdf	P value	CFI	Δ CFI	RMSEA	Δ RMSEA	RMSEA 90 % CI
Configural invariance		848.854	582			<.001	0.98	0.00	0.032		0.027–0.037
Metric invariance	Measurement	928.323	647	79.47	65	<.001	0.98	0.00	0.032	0.000	0.027–0.037
	Structural	948.323	667	20.00	20	<.001	0.98	0.00	0.032	0.000	0.027–0.034
Intercept invariance	Measurement	1361.692	722	413.37	55	<.001	0.96	0.02	0.044	0.012	0.041–0.048
	Structural	1644.448	747	282.756	25	<.001	0.95	0.01	0.051	0.007	0.048–0.056

Data and table copyright Pearson. Comparison ($N = 2637$): Caucasian females $n = 830$, Caucasian males $n = 777$, Africa–American females $n = 221$, African–American males $n = 188$, Hispanic females $n = 313$, and Hispanic males $n = 308$)

CFI comparative fit index, RMSEA root-mean-square error of approximation

intercept invariance of both the first- and second-order factor loadings could be concluded.

Conclusion In sum, configural, metric, and intercept invariance were established for every CHC factor and *g* across all three ethnic/gender group comparisons. Although, as expected, the added constraints resulted in significant values in χ^2 , the CFI and RMSEA values were more defensible statistics for evaluating model fit, given the sample size, the number of constraints, and the general complexity of the model. Based on the very good CFI and RMSEA values and the non-significant or small changes observed in CFI and RMSEA values as more constraints were added, it was concluded that configural, metric, and intercept invariance were established for all six groups for each CHC factor and for *g*.

Discussion

Overall, results show a consistent factor structure of the WISC-V across African–American, Hispanic, and Caucasian male and female students from age 6 to 16. Even though the mean scores between the groups differ, sometimes substantially, fit indexes suggest structural equivalence for all five broad CHC factors and for *g*. Importantly, the structural invariance found in this study indicates (a) that the scores are equally meaningful for students of African–American, Caucasian, and Hispanic origin and (b) that equality extends to separate groups of males and females within each of these three ethnic groups. Such results are consistent with previous research findings that suggest factorial invariance for different ethnic or gender groups separately across other intelligence tests, including the Kaufman scales (Reynolds et al. 2015; Scheiber 2016a), the Woodcock–Johnson tests (Edwards & Oakland, 2006; Keith 1999), and the Differential Ability Scales (Keith et al. 1999; Trundt 2013). This is the first study to have established factorial invariance across different ethnic/gender groups together and the first study to provide evidence for structural equivalence by gender or ethnicity of the WISC-V.

Overall, results suggest that inferences drawn from the WISC-V scores are similarly meaningful for African–American, Hispanic, and Caucasian male and female students ages 6 to 16. This satisfies APA standards 7.1 and 7.8 for the fair use of the test with these subgroups in terms of its psychometric validity. However, this is not to say that the mean score differences observed are valid indicators of “true” differences in ability. Discrepancies in mean scores may occur for a variety of reasons, most of which relate to differences in socioeconomic status and home environment (e.g., Kaufman et al. 2016; Weiss et al. 2006, 2015), but have little to do with the test’s psychometric properties and do *not* provide evidence that any particular ethnic/gender group is “smarter” than any other group.

Clinical Implications

Across the country, school psychologists, neuropsychologists, and clinical psychologists administer cognitive tests of intelligence to children of different ethnic/gender groups based on the assumption that those tests are valid and reliable. However, with the exception of a few research studies that have investigated construct invariance of cognitive tests across different ethnic or gender groups separately (e.g., Edwards & Oakland, 2006; Keith 1999; Reynolds et al. 2015; Scheiber 2016a; Trundt 2013), there is a scarcity of empirical evidence to support the assumption of construct invariance across different gender/ethnic groups together in individually administered tests of cognition. This is the first construct validity study of the WISC-V conducted with ethnic/gender groups together. Results of the study provide further empirical evidence for the continued use of the WISC-V with those subgroups.

Results of this study are especially important as the WISC-V is the most recent version of the Wechsler scales and, thus, is—and will be for several years to come—the most commonly used intelligence test all across the country to determine diagnosis and access or denial to special programs. Based on the results of this study, school psychologists and neuropsychologists can have increased confidence that the WISC-V test scores, when interpreted based on CHC theory, measure what they are supposed to measure for African–American, Hispanic, and Caucasian male and female children ages 6 to 16.

Furthermore, the present results are consistent with Carroll’s (1993) assertion that CHC theory is a universal theory of intelligence and applicable to different gender and ethnic groups. Results of the current study provide evidence that the five Stratum II broad ability factors in addition to the Stratum III factor (*g*) are construct invariant and, thus, have comparable meanings for students across different ethnic gender groups. Mean score difference studies have demonstrated different mean scores in terms of magnitude on the WISC-V for different ethnic and gender groups (e.g., Kaufman et al. 2016; Wechsler 2014b; Weiss et al. 2015). Thus, it was important to investigate construct invariance of each single CHC factor separately in addition to *g*. The inclusion of different ethnic/gender groups was an especially unique contribution of this study. Given that construct invariance of the broad ability factors and *g* was established across different ethnic/gender groups provides even more evidence for the generalizability of CHC theory to various subgroups.

Limitations

As with every study, data need to be understood in view of the study’s limitations: perhaps most importantly, the sample was matched on 2012 US Census data in terms of important

background variables, including age, gender, ethnicity, geographic region, and parental educational attainment (as a proxy for socioeconomic status). Parent educational attainment was not controlled, as the aim of this study was to evaluate ethnic bias based on how cognitive tests are actually used in society. However, future studies should replicate analyses of the current study across different socioeconomic levels by splitting their groups by parental education or other socioeconomic variables to evaluate whether results maintain for different SES groups. Furthermore, the focus of the present study was to investigate construct invariance across the different CHC latent factors and *g*; however, future research should conduct differential item functioning analysis in order to assess construct invariance on the subtest level. Furthermore, this study only focused on the five core CHC factors measured by the WISC-V, including *Gc*, *Gs*, *Gf*, *Gsm*, and *Gp*; however, future researchers should also validate the sixth CHC factor measured by the WISC-V (*Glr*). In addition, although construct invariance was validated for African–American, Hispanic, and Caucasian males and female students ages 6–16, such results are not generalizable to other ethnic or age groups. Future research should also replicate results with other individually administered tests of cognition. Finally, it is important to take into consideration that the sample consisted primarily of normally developing children, with proportional representation of exceptional children. However, those children that are most commonly referred to psychological testing are the precise children and adolescents who were underrepresented in this study. Future researchers should conduct studies with large populations of students who struggle with learning or have intellectual disabilities.

Conclusions

Overall, results of the current study provide evidence for the construct invariance of the most recent version of the Wechsler tests—the WISC-V (Wechsler 2014a)—across six different ethnic/gender groups. Findings suggest that the factor structure of the WISC-V based on CHC theory was invariant across a representative sample of African–American, Hispanic, and Caucasian male and female students ages 6–16. Findings provide evidence for continued use of the WISC-V when it comes to the assessment of intellectual ability for African–American, Hispanic, and Caucasian students, both male and female. However, it is important to emphasize that even though the WISC-V might not be biased in terms of its constructs, the mean differences in test scores require further investigation. Cognitive tests are impacted by a number of different factors (including income, home environment, quality of schooling, cultural opportunities) that are impossible to control for but have been found to influence cognitive test scores significantly (Weiss et al. 2006, 2015). Thus, it is

essential not to draw any conclusions based on mean score differences found across different subgroups on the WISC-V or any other cognitive test instruments.

Compliance with Ethical Standards All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000 (5). Informed consent was obtained from all participants for being included in the study, as per WISC-V Technical and Interpretative Manual (Wechsler 2014b).

Conflict of Interest The author declares that she has no conflict of interest.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Arbuckle, James L. (1995–2011). *AMOS 20.0 User's guide*. Crawfordville, FL: AMOS Development Corporation.
- Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, 107, 238–246.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456–466. doi:10.1037/0033-2909.105.3.456.
- Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies*. New York: Cambridge University Press.
- Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: testing measurement invariance of second-order factor models. *Structural Equation Modeling*, 12(3), 471–492.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255. doi:10.1207/S15328007SEM0902_5.
- Crocker, J., Luhtanen, R. K., & Cooper, M. L. (2003). Contingencies of self-worth in college students: theory and measurement. *Journal of Personality and Social Psychology*, 85, 894–908.
- Edwards, O. W., & Oakland, T. D. (2006). Factorial invariance of Woodcock-Johnson III scores for African Americans and Caucasian Americans. *Journal of Psychoeducational Assessment*, 24(4), 358–366. doi:10.1177/0734282906289595.
- Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Dynda, A. (2014). Cognitive assessment: progress in psychometric theories, the structure of cognitive tests, and approaches to test interpretation. In D. Saklofske, V. Schwane, & C. Reynolds (Eds.), *Oxford handbook of psychological assessment of children and adolescents*. New York: Oxford University Press.
- Jirout, J. J., & Newcombe, N. S. (2015). Building blocks for developing spatial skills evidence from a large, representative US sample. *Psychological Science*, 0956797614563338.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman assessment battery for children—second edition*. Circle Pines: American Guidance Service.
- Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligence testing with the WISC-V*. Hoboken: Wiley.
- Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: similarities and differences across ethnic groups. *School Psychology Quarterly*, 14(3), 239.

- Keith, T. Z., Quirk, K. J., Scharzter, C., & Elliott, C. D. (1999). Construct bias in the differential ability scales? Confirmatory and hierarchical factor structure across three ethnic groups. *Journal of Psychoeducational Assessment*, 17(3), 249–268. doi:[10.1177/073428299901700305](https://doi.org/10.1177/073428299901700305).
- 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-: What does it measure?. *School Psychology Review*, 35(1), 108.
- Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence*, 36(6), 502–525.
- Keith, T. Z., Reynolds, M. R., Roberts, L. G., Winter, A. L., & Austin, C. A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: evidence from the Differential Ability Scales—Second Edition. *Intelligence*, 39(5), 389–404.
- Lichtenberger, E. O., & Kaufman, A. S. (2013). *Essentials of WAIS-IV assessment* (2nd ed.). Hoboken: Wiley.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543.
- Meyers, L.S., Gamst, G., & Guarino, A.J. (2013). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: SAGE Publications Inc.
- Palejwala, M. H., & Fine, J. G. (2014). Gender differences in latent cognitive abilities in children aged 2 to 7. *Intelligence*, 48, 96–108.
- Reynolds, M. R., & Keith, T. Z. (2013). Measurement and statistical issues in child assessment research. In C. R. Reynolds (Ed.), *Oxford handbook of psychological assessment of children and adolescents*. New York: Oxford University.
- Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 332–374). Hoboken: Wiley.
- Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, 36(3), 236–260.
- Reynolds, M. R., Scheiber, C., Hajovsky, D. B., Schwartz, B., & Kaufman, A. S. (2015). Gender differences in academic achievement: is writing an exception to the gender similarities hypothesis? *Journal of Genetic Psychology*, 176(4), 211–234.
- Scheiber, C. (2016a). Does the KABC-II display ethnic bias in the prediction of reading, math, and writing in elementary school through high school? *Assessment*.
- Scheiber, C. (2016b). Do the Kaufman tests of cognitive ability and academic achievement display construct bias across a representative sample of Black, Hispanic, and Caucasian school-age children in grade 1 through 12? *Psychological Assessment*.
- Scheiber, C., & Kaufman, A. S. (2015). Which of the three KABC-II global scores is the least biased? *Journal of Pediatric Neuropsychology*, 1(4), 21–35.
- Scheiber, C., Reynolds, M. R., Hajovsky, D., & Kaufman, A. S. (2015). Gender differences in achievement in a large, nationally-representative sample of children and adolescents. *Psychology in the Schools*, 52(4), 335–348.
- Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. *Contemporary intellectual assessment: theories, tests, and issues* (3rd), 99–144.
- Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling: A Multidisciplinary Journal*, 5(4), 411–419. doi:[10.1080/10705519809540115](https://doi.org/10.1080/10705519809540115).
- Trundt, K. M. (2013). *Construct bias in the differential ability scales, (DAS-II): a comparison among African American, Asian, Hispanic, and White ethnic groups* (Published doctoral dissertation). Austin, TX: University of Texas.
- Van der Vegt, G. S., Emans, B. J. M., & Van der Vliet, E. (2001). Patterns of interdependence in work teams: a two-level investigation of the relations with job and team satisfaction. *Personnel Psychology*, 54, 51–69.
- Wechsler, D. (2014a). *Wechsler intelligence scale for children* (5th ed.). Bloomington: PsychCooperation.
- Wechsler (2014b). WISC-V: Wechsler intelligence scale for children-fifth edition. Technical and interpretive manual. San Antonio, TX: PsychCorp.
- Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (2006). *WISC-IV advanced clinical interpretation*. Burlington, MA: Academic.
- Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). Technical and practical issues in the structure and clinical invariance of the Wechsler scales: A rejoinder to commentaries. *Journal of Psychoeducational Assessment*, doi:[10.1177/0734282913478050](https://doi.org/10.1177/0734282913478050).
- Weiss, L. G., Saklofske, D. H., Holdnack, J. A., & Prifitera, A. (2015). *WISC-V assessment and interpretation*. Waltham, MA.
- Wendling, B. J., & Mather, N. (2009). *Essentials of evidence-based academic interventions*. Hoboken: Wiley.
- Williams, R. L. (1971). Abuses and misuses in testing Black children. *The Counseling Psychologist*, 2(3), 62–73. doi:[10.1177/001100007100200314](https://doi.org/10.1177/001100007100200314).