


Measurement Invariance of the Wechsler Intelligence Scale for Children, Fifth Edition 10-Subtest Primary Battery: Can Index Scores be Compared across Age, Sex, and Diagnostic Groups?

Journal of Psychoeducational Assessment
2021, Vol. 39(1) 89–99
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0734282920954583
journals.sagepub.com/home/jpa


Stefan C. Dombrowski¹ , Marley W. Watkins², Ryan J. McGill³, Gary L. Canivez⁴ , Calliope Holingue⁵, Alison E. Pritchard⁵, and Lisa A. Jacobson⁵

Abstract

Measurement invariance of the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V) 10 subtest primary battery was evaluated across sex, age (6–8, 9–11, 12–14, and 15–16 year-olds), and three diagnostic (attention-deficit/hyperactivity disorder, anxiety, and encephalopathy) groups within a large clinical sample ($N = 5359$) referred to a children's specialty hospital. Competing models were tested using confirmatory factor analysis (CFA), and a five-factor oblique model corresponding to the publisher's hypothesized first-order measurement model (e.g., verbal comprehension, fluid reasoning, visual-spatial, working memory, and processing speed) was found to have the best model fit. Multigroup CFA was subsequently used to evaluate progressively more restrictive constraints on the measurement model. Results indicated that full metric invariance was attained across the three groups studied. Full scalar invariance was attained for sex and diagnostic groups. Partial scalar invariance was attained for age-group. The results of this study provide support for the first-order scoring structure of the five WISC-V factors in the 10 subtest primary battery with this large clinical sample.

Keywords

Wechsler intelligence scale for children, fifth edition, measurement invariance, multigroup confirmatory factor analysis

¹Rider University, NJ, USA

²Baylor University, TX, USA

³William & Mary, VA, USA

⁴Eastern Illinois University, IL, USA

⁵Kennedy Krieger Institute, Johns Hopkins University School of Medicine, MD, USA

Corresponding Author:

Stefan C. Dombrowski, School Psychology Program, Rider University, 2083 Lawrenceville Road, Lawrenceville 08648, NJ, USA.

Email: sdombrowski@rider.edu

Measurement invariance is an important but often underutilized aspect of construct validation for cognitive ability instruments. It assists in determining whether the measurement model holds across different groups nested within a broader target population being studied and, ultimately, whether resulting index/composite scores can be confidently interpreted across groups in the same way, if at all (Meredith, 1993). Measurement invariance may be conceptualized as a more technical extension of structural validity. Instead of focusing on a single group as occurs in traditional exploratory and confirmatory factor analysis (CFA) (see, e.g., Dombrowski et al., 2018a, 2018b), it works by constraining selected parameters to equality, permitting a “stress test” on various elements of an instrument’s structure across the groups studied with the ultimate goal of determining whether an instrument’s scores may be compared among those groups. The most common way to assess measurement invariance is through multigroup confirmatory factor analysis (MGCFA; Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014). In MGCFA, models are fitted to data using different sets of constraints corresponding to different levels or types of invariance. Researchers typically differentiate among three levels of measurement invariance that are sufficient for conducting most comparative data analyses: configural, metric, and scalar invariance (Vandenberg & Lance, 2000; see Meredith, 1993 for additional, more restrictive levels that are not commonly applied).

After a plausible baseline measurement model is identified, the first step in MGCFA is to determine whether the pattern of the loadings is the same across groups (configural invariance). Once these constraints are applied and no meaningful attenuation in representative fit statistics is observed, it can be reasonably concluded that the test has equal form (or configuration) across groups. Once equal form is established, the equivalence of the magnitude of the loadings is evaluated (metric invariance). Within the assessment literature, metric invariance is frequently referred to as a form of *weak* invariance (Putnick & Bornstein, 2016). Finally, the latent intercepts are evaluated to determine whether they are equivalent across groups (scalar invariance). If so, then an instrument is thought to have attained *strong* invariance (Putnick & Bornstein, 2016), and the factor scores may be confidently interpreted across groups.

Invariance analyses have been conducted on the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V; Wechsler, 2014a) using the 16 primary and secondary subtest normative sample data to determine equivalence of age and sex (sic., gender) in both US standardization samples (e.g., Reynolds & Keith, 2017; Scheiber, 2016) and international samples (e.g., Chen, Zhu, Liao, & Keith, 2020; Pauls, Daseking, & Petermann, 2019). However, only one invariance study has been conducted on the 10-primary subtest battery. Specifically, Graves, Smith, and Nichols (2020) investigated the invariance of the 10-primary subtest battery in a predominantly African American sample and found that the structure failed to attain full metric invariance. The lack of analyses on the 10-primary subtest battery is noteworthy considering that it contains the most frequently administered group of subtests by practitioners (Benson et al., 2019). Additionally, analyses on referred clinical samples (e.g., attention-deficit/hyperactivity disorder (ADHD), anxiety, and brain injury) are important but rarely investigated (Chen et al., 2020). Children referred for the evaluation of suspected disability are the ones most frequently administered tests of cognitive ability, and there are frequent calls for analyses with clinical samples (Chen et al., 2020; Graves et al., 2020). However, such analyses are less available in the literature. Accordingly, the purpose of this study was to investigate the measurement invariance of the 10 primary WISC-V subtests across sex (male/female), diagnostic group (ADHD, anxiety, and encephalopathy¹), and four age groups (6–8, 9–11, 12–14, and 15–16 year-olds) with a large clinical sample.

Method and Data Analyses

A total of 5359 children between the ages of 6 and 16 years were administered the 10 WISC-V primary subtests as part of clinical assessments completed at a large outpatient clinical psychology/neuropsychology clinic in a children’s specialty hospital. De-identified data from

completed clinical evaluations were retrieved from the electronic medical record database following approval by the hospital's Institutional Review Board as well as representative university equivalents. The sample was primarily composed of White/Caucasian and Black/African American youth (Table 1). Three diagnostic groups (ADHD, 47.6%; anxiety, 11.6%; and encephalopathy, e.g., nontraumatic diffuse brain dysfunction; 10.1%) comprised just over two-thirds of the sample (Table 2). The participants' ages ranged from 6.0 to 16.93 years ($M = 10.69$, $SD = 2.74$). Compared with the US standardization sample, this sample was slightly below average in subtest and composite scores, as is typical in clinical samples (Table 3). All subtest and composite scores showed univariate normal distributions with no appreciable skewness or kurtosis. However, Mardia (1970) multivariate estimates for the sample (skewness = 1.344; kurtosis = 125.128) indicated significant ($p < .0001$) multivariate nonnormality (Cain, Zhang, & Yuan, 2017).

Mplus version 8.4 (Muthén & Muthén, 1998–2017) was used to estimate all baseline comparison and multiple group CFA models. The robust maximum likelihood estimator (i.e., Satorra & Bentler, 1988) that adjusts for nonnormality was used to calculate estimates and corrected fit indices. The first-order five-factor measurement model (nested within the higher order Model 5e, Figure 5.2, Wechsler, 2014b, p. 84) for the WISC-V was tested against competing models (see Table 4). Once a baseline model was established, it was separately fit to each group to ensure model adequacy. Subsequently, a series of progressively restrictive constraints was applied on sets of model parameters to determine equivalence across sex (male/female), diagnostic group (ADHD, anxiety, and encephalopathy), and age-group (6–8, 9–11, 12–14, and 15–16 year-olds). Initially, the equivalence of the WISC-V across the instrument's structure (configural invariance) was investigated. No between-group parameter constraints were imposed other than those fixed to a value (i.e., setting the scale [1.0]) to adequately identify the model. Between-group constraints were then imposed on all factor loadings except for the loadings on the referent indicator in each group (metric invariance), which were set to 1.0. In scalar specification, the intercepts were constrained to equality, while the factor mean was fixed to zero in the first group but was allowed to be freely estimated in other groups. Differences to reject invariance across all specifications were $\Delta CFI \leq .01$ and $\Delta RMSEA \leq .015$ for both factor loadings and intercepts and $\Delta SRMR \leq .03$ for factor loadings and $\leq .01$ for intercepts (Chen, 2007; Cheung & Rensvold, 2002).

Results

The comparison of baseline models (Table 4) suggested that an oblique five-factor first-order model, corresponding to the first-order WISC-V measurement structure (see Figure 1), provided the best statistical fit ($S-B\chi^2(25) = 222.2$, $p < .05$; comparative fit index (CFI) = .993) to these data.

Table 1. Demographic Characteristics of the Clinical Sample.

Race/Ethnicity	N	Percent	Sex		
			Female	Male	Unknown
White	2865	53.50	885	1764	216
Black	1513	28.20	443	940	130
Multiracial	376	7.00	117	257	2
Unknown/Other	209	3.90	40	100	69
Asian	191	3.60	54	108	29
Hispanic/Spanish origin	190	3.50	57	132	1
American Indian/Alaskan native	12	.20	5	7	0
Native Hawaiian or Pacific Islander	3	.10	1	0	1
Total	5359		1602	3308	448
Percent		100.00	29.90	61.70	8.40

Table 2. Diagnostic Categories of the Clinical Sample.

ICD diagnosis	N	Percent
ADHD/ADD	2552	47.6
Encephalopathy	620	11.6
Anxiety	539	10.1
Adjustment disorder	213	4.0
Behavior disorder	213	4.0
Epilepsy	140	2.6
Mood disorder	133	2.5
Congenital anomaly	112	2.1
Genetic condition	112	2.1
Frontal lobe deficit	100	1.9
Disorder of the nervous system	76	1.4
Major depression	69	1.3
Brain/spine injury	43	.8
Neoplasm/Tumor	40	.7
Hearing loss	38	.7
Leukemia	38	.7
Unknown	37	.7
Other depressive disorder	30	.6
Autism spectrum disorder	25	.5
Cancer (not brain/nervous system)	23	.4
Emotional disturbance	16	.3
Expressive/receptive language disorder	13	.2
Fetal alcohol syndrome	13	.2
Bipolar disorder	11	.2
Other mental/psychological disorder	11	.2
Reading/learning disability	11	.2
Tic/Tourette's disorder	11	.2
Misc medical/psychiatric conditions	120	2.2
Total	5359	100.0

Note. ICD = international classification of diseases, tenth edition; ADD = attention deficit disorder; ADHD = attention-deficit/hyperactivity disorder

In addition to having the best model fit among the competing models, the oblique five-factor model was selected on the basis that it is the model that guided how the test publisher recommends the instrument is scored and interpreted and subsequently how the majority of psychologists *actually* interpret the test in practice (Benson et al., 2019). There is another essential point. Although the publisher preferred theoretical model is a conventional higher order model where the influence of general intelligence on the measured variables is fully mediated through the first-order group factors, the scoring structure of the WISC-V primary subtests does not reflect this model: only seven of 10 subtests load the general factor, while all 10 subtests contribute to their respective group factors. Since a general intelligence factor is fully mediated through group factors, this model cannot be readily evaluated using higher order factor analysis. Thus, the test publisher never fully examined its promoted scoring structure for the WISC-V 10-primary subtest battery, nor has it ever been evaluated in the extant literature until this study. A bifactor approach can model this structure albeit in a slightly different form than the bifactor models that have been previously investigated in the literature (see Table 4, “bifactor 5 score” for fit indices results).

Table 3. Descriptive Statistics for the WISC-V Clinical Sample.

Subtest/Composite	Total sample (N = 5359)			
	M	SD	Skewness	Kurtosis
Subtests				
Block design	8.69	3.33	.12	-.17
Similarities	9.14	3.30	-.02	-.07
Matrix reasoning	9.02	3.38	.08	-.10
Digit span	7.96	3.10	.11	.07
Coding	7.57	3.33	-.03	-.37
Vocabulary	8.99	3.56	.06	-.53
Figure weights	9.52	3.12	-.03	-.28
Visual puzzles	9.58	3.29	-.04	-.43
Picture span	8.55	3.12	.14	-.16
Symbol search	8.21	3.23	.00	.03
Composites				
VCI	94.96	17.41	-.04	-.18
VSI	95.15	17.23	.08	-.06
FRI	95.80	16.77	.02	-.37
WMI	89.94	15.82	.14	-.13
PSI	88.10	17.10	-.15	.03
FSIQ	91.03	17.27	-.00	.04

Note. VCI = verbal comprehension index; VSI = visual spatial index; FRI = fluid reasoning index; WMI = working memory index; PSI = processing speed index; WISC-V = Wechsler intelligence scale for children, fifth edition; FSIQ = full scale intelligence quotient.

The fit of the confirmatory factor analytic models for the sex, age, and diagnostic groups demonstrated that all single-group models fit roughly equivalently to the total sample model (see Table 4). Model fit indices and statistics for progressively constrained models are presented in Table 5. The configural model fits these data well across sex, age, and diagnostic group with no appreciable loss in model fit. When the more restrictive metric constraints were applied, there was also no loss in discernable model fit across all three groups that were evaluated ([sex] Δ CFI = .000; Δ SRMR = .002; Δ RMSEA = .001; Δ AIC = 000; [age] Δ CFI = .005; Δ SRMR = .019; Δ RMSEA = .008; Δ AIC = -129; [diagnostic group] Δ CFI = .000; Δ SRMR = -.002; Δ RMSEA = .000; Δ AIC = 001). Constraining intercepts to equality indicated that the diagnostic group (Δ CFI = .005; Δ SRMR = .019; Δ RMSEA = .008) and the sex group (Δ CFI = .005; Δ SRMR = .019; Δ RMSEA = .008) met the threshold for scalar invariance. The age group (Δ CFI = .014; Δ SRMR = .024; Δ RMSEA = .020) did not. Inspection of modification indices suggested that the fluid reasoning subtest intercepts were problematic causing the lack of full scalar invariance. This was similarly found in a study that investigated invariance of the WISC-V in referred Black/White sample (Graves et al., 2020). Partial scalar invariance was subsequently achieved by freeing the figure weights intercept which significantly improved overall model fit (see Table 5).

Discussion

The investigation of measurement invariance requires consideration of whether an instrument's factor structure, factor loadings, and intercepts are equivalent across groups when subjected to increasingly restrictive parameter constraints. The evaluation of invariance provides an opportunity to impose a psychometric stress test on the structure of an instrument (thereby further substantiating its structural validity beyond single-group analyses, if attained). With scalar

Table 4. Model Fit of Competing Models and Separate Groups.

Model	S-By ²	df	CFI	ΔCFI	TLI	ΔTLI	SRMR	ΔSRMR	RMSEA	90% CI RMSEA	ΔRMSEA	AIC	ΔAIC	BIC	ΔBIC
Total sample (N = 5359)															
Oblique 5-factor	222.2	25	.993	—	.987	—	.013	—	.038	.034–.043	—	250034	—	250298	—
Oblique 4-factor	405.4	29	.987	.006	.979	.008	.018	.005	.049	.045–.054	.011	250218	184	250455	157
Higher order 5-factor	836.6	30	.972	.021	.957	.030	.033	.020	.071	.069–.077	.033	250668	634	250899	601
Higher order 4-factor	570.8	31	.981	.012	.972	.015	.025	.012	.057	.053–.061	.019	250386	352	250610	312
Bifactor 4 factor	425.4	28	.986	.007	.977	.010	.022	.009	.052	.047–.056	.018	250240	206	250484	186
Bifactor 5-factor score	6773.2	31	.762	.231	.655	.332	.293	.28	.202	.198–.207	.164	256651	6617	256875	6577
Bifactor 4-factor score	6688.2	31	.765	.228	.659	.328	.291	.28	.200	.196–.204	.162	256566	6532	256790	6492
Baseline—Oblique 5-factor male (n = 3307) and female (n = 1606)															
Oblique 5-factor male	143.2	25	.993	—	.988	—	.013	—	.038	.032–.044	—	155160	—	155404	—
Oblique 5-factor female	84.5	25	.994	—	.988	—	.015	—	.038	.030–.048	—	74128	—	74343	—
Baseline—Oblique 5-factor age 6 to 8 (n = 1735), age 9 to 11 (n = 1818), age 12 to 14 (n = 1367), and age 15 + (n = 405)															
Oblique 5 age 6–8	97.7	25	.991	—	.983	—	.018	—	.041	.033–.050	—	81949	—	82167	—
Oblique 5 age 9–11	101.3	25	.993	—	.987	—	.014	—	.041	.033–.049	—	84684	—	84904	—
Oblique 5 age 12–14	90.6	25	.992	—	.985	—	.014	—	.044	.034–.054	—	62630	—	62839	—
Oblique 5 age 15–16	34.5	25	.996	—	.994	—	.017	—	.031	.000–.054	—	18740	—	18900	—
Baseline oblique 5 factor—ADHD (n = 2277), anxiety (n = 484), and encephalopathy (n = 541)															
Oblique 5-factor ADHD	86.8	25	.994	—	.990	—	.013	—	.033	.026–.041	—	105987	—	106216	—
Oblique 5-factor anxiety	21.5	25	1.00	—	1.000	—	.011	—	.000	.000–.030	—	22451	—	22618	—
Oblique 5-factor enceph.	63.1	25	.988	—	.979	—	.020	—	.053	.037–.070	—	25826	—	25998	—

Note. Best index fit in bold. Δ in comparison to best fit in total sample. Fit of higher order and bifactor five-factor models identical due to constraints imposed to identify bifactor model. The small number of indicators per factor makes it necessary to constrain loadings which, in turn, makes invariance tests of bifactor models not of the actual data but of a constrained version of the data. Bifactor 5-factor score = actual implied scoring structure of WISC-V where 7 subtests load g and 10 load the respective 5 factors. Bifactor 4 score = same as BF 5 score except fluid reasoning and visual-spatial subtests fused to form perceptual reasoning factor. Enceph. = encephalopathy; WISC-V = Wechsler intelligence scale for children, fifth edition. CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; AIC = Akaike information criterion; BIC = Bayesian information criterion.

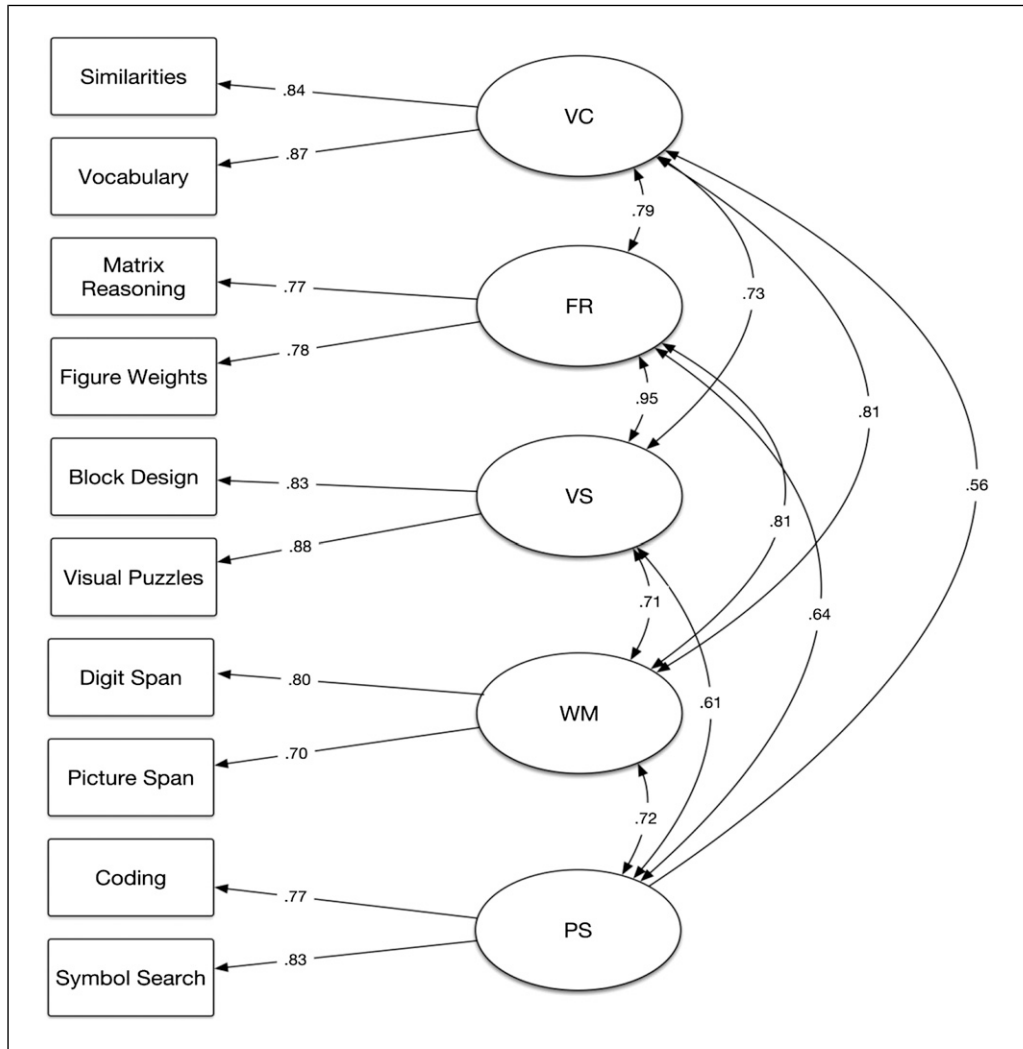


Figure 1. Wechsler Intelligence Scale for Children, Fifth Edition baseline measurement model identified by the present clinical sample.

invariance specification, where constraints are placed on intercepts, invariance determines whether an instrument’s scores (in this case, the WISC-V index scores) may be compared and whether resulting scores are not confounded by an artifact of the measurement structure.

The consideration of invariance is important to aid in better understanding the WISC-V 10-subtest primary battery. It has been argued that the WISC-V primary battery theoretical structure was essentially extrapolated from the 16-subtest battery (Dombrowski, Canivez, & Watkins, 2017). Thus, important information regarding the structure of the 10-subtest WISC-V is less available.

The present study evaluated numerous competing models and found that the five-factor oblique model, which also reflects the instrument’s scoring structure, had the best model fit (see Table 4) with this clinical sample. Invariance testing proceeded using this model as baseline. The results suggested that the WISC-V evidenced weak (metric) invariance across all groups (sex [male/female],

Table 5. Invariance of the Five Oblique WISC-5 Measurement Model—Gender, Age, and Diagnostic Group Samples.

Model	S- $B\chi^2$	df	CFI	Δ CFI	TLI	Δ TLI	SRMR	Δ SRMR	RMSEA	90% CI RMSEA	Δ RMSEA	AIC	Δ AIC	BIC	Δ BIC
Invariance—Male ($n = 3307$) versus female ($n = 1606$)															
Configural	237.6	50	.993	—	.987	—	.014	—	.039	.034–.044	—	229287	0	229807	—
Metric	246.8	55	.993	.000	.988	.001	.016	.02	.038	.033–.043	.001	229287	0	229774	-33
Scalar*	345.9	60	.989	.004	.984	.004	.017	.03	.044	.040–.049	.006	229376	89	229831	24
Invariance—Age 6 to 8 ($n = 1742$), 9 to 11 ($n = 1823$), 12 to 14 ($n = 1378$), and 15 to 16 ($n = 405$)															
Configural	327.5	100	.992	—	.986	—	.016	—	.041	.036–.046	—	249081	—	250134	—
Metric	490.2	115	.987	.005	.980	.006	.035	.019	.049	.045–.054	.008	249210	-129	250165	34
Scalar**	768.4	180	.978	.014	.969	.017	.040	.024	.061	.056–.065	.020	249462	-381	250318	184
Part scalar ^a	592.2	127	.984	-.003	.977	-.003	.036	.001	.052	.048–.057	.003	249290	80	250165	0
Invariance—ADHD ($n = 2277$), anxiety ($n = 484$), and encephalopathy ($n = 541$)															
Configural	170.9	75	.994	—	.990	—	.016	—	.034	.027–.041	—	154264	—	154996	—
Metric	194.9	85	.994	.000	.990	.000	.014	-.002	.034	.027–.041	.000	154265	1	154936	-60
Scalar***	271.9	95	.990	.004	.985	.005	.023	.009	.041	.035–.047	.007	154324	59	154934	-2

Note. * Scalar versus metric S- $B\Delta\chi^2$ (5) = 99.1, $p < .0001$. ** Scalar versus metric S- $B\Delta\chi^2$ (15) = 279.2, $p < .0001$. *** Scalar versus metric S- $B\Delta\chi^2$ (10) = 78.42, $p < .0001$. WISC-V = Wechsler intelligence scale for children, fifth edition. CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; AIC = Akaike information criterion; BIC = Bayesian information criterion.

^aFigure weights intercept freely estimated.

age [6–8, 9–11, 12–14, and 15–16], and diagnostic group [ADHD, anxiety, and encephalopathy]) investigated. An evaluation of the strong (scalar) specification suggested that both sex and diagnostic groups attained full scalar invariance, while age-groups attained partial scalar invariance. The finding of full or partial scalar invariance was consistent with previous research findings with the extended WISC-V 16 primary and secondary subtest battery (e.g., [Chen et al., 2020](#); [Pauls et al., 2019](#); [Reynolds & Keith, 2017](#); [Scheiber, 2016](#)) and other intelligence tests including the Kaufman assessment battery for children, second edition ([Reynolds, Scheiber, Hajovsky, Schwartz, & Kaufman, 2015](#); [Scheiber, 2017](#)), Woodcock-Johnson ([Edwards & Oakland, 2006](#); [Keith, 1999](#)), and differential ability scale ([Keith, Quirk, Scharzter, & Elliott, 1999](#)). However, this is the first study to investigate invariance of the WISC-V 10-primary subtest battery across several groups (e.g., sex, age, and clinical diagnosis) with a referred sample more than double the size of the normative sample. The present study's conclusions are limited by a lack of comparison to the standardization sample². This would have offered another vantage from which to assess invariance.

Conclusion and Implications for Practice

The present results have implications for interpretation of the broader measurement model in clinical practice and suggest that individuals from different sex, age, and clinical groups may have their index scores confidently compared to one another ([Rudnev, Lytkina, Davidov, Schmidt, & Zick, 2018](#)). This conclusion is particularly important for the clinical comparison group. There have been recent calls for structural validity and invariance analyses within clinical groups ([Chen et al., 2020](#); [Graves et al., 2020](#)), but rarely are data sets such as the one in the present study available. In this case, although three distinctly different clinical groups were available—an externalizing disorder (i.e., ADHD), an internalizing disorder (i.e., anxiety), and a neurologically based disorder (i.e., encephalopathy)—the scoring structure was either fully or partially invariant and functions the same way regardless of sex, age, or clinical group. Stated another way, users of the WISC-V 10-primary subtest battery can be more confident that a score obtained on one of the WISC-V indices is a function of performance by a group member and unrelated to statistical distortions in the measurement instrument due to age, sex, or clinical condition. In sum, the attainment of configural, metric, and scalar invariance across three different groups with this clinical sample lends evidentiary support for the viability of the WISC-V first-order factors in clinical practice and suggests that the 10 WISC-V primary subtest battery measures intended first-order constructs with this sample in the way proposed by the publisher.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Stefan C. Dombrowski  <https://orcid.org/0000-0002-8057-3751>

Gary L. Canivez  <https://orcid.org/0000-0002-5347-6534>

Notes

1. Encephalopathy is a generic category that includes brain injury of a diffuse but nontraumatic nature.
2. Our request for the standardization sample data was denied by NCS Pearson, Inc. Invariance analyses, like the present study, are not included in the WISC-V *Technical and Interpretive Manual*.

References

- Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 national survey. *Journal of School Psychology, 72*, 29-48. doi:10.1016/j.jsp.2018.12.004
- Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods, 49*, 1716-1735. doi:10.3758/s13428-016-0814-1
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 464-504. doi:10.1080/10705510701301834
- Chen, H., Zhu, J., Liao, Y.-K., & Keith, T. Z. (2020). Age and gender invariance in the Taiwan Wechsler intelligence scale for children, fifth edition: Higher order five-factor model. *Journal of Psychoeducational Assessment, 17*. doi:10.1177/0734282920930542
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255. doi:10.1207/S15328007SEM0902_5
- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55-75. doi:10.1146/annurev-soc-071913-043137
- Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2017). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology, 22*, 90-104. doi:10.1007/s40688-017-0125-2
- Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018a). An alternative conceptualization of the theoretical structure of the Woodcock-Johnson IV tests of cognitive abilities at school age: A confirmatory factor analytic investigation. *Archives of Scientific Psychology, 6*, 1-13. doi:10.1037/arc0000039
- Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018b). Hierarchical exploratory factor analyses of the Woodcock-Johnson IV full test battery: Implications for CHC application in school psychology. *School Psychology Quarterly, 33*, 235-250. doi:10.1037/spq0000221
- 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*, 358-366. doi:10.1177/0734282906289595
- Graves, S. L., Smith, L. V., & Nichols, K. D. (2020). Is the WISC-V a fair test for black children: Factor structure in an urban public school sample. *Contemp School Psychol*. doi:10.1007/s40688-020-00306-9
- Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly, 14*, 239-262. doi:10.1037/h0089008
- Keith, T. Z., Quirk, K. J., Schartzler, 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*, 249-268. doi:10.1177/073428299901700305
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519-530. doi:10.1093/biomet/57.3.519
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543. doi:10.1007/bf02294825
- Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.
- Pauls, F., Daseking, M., & Petermann, F. (2019). Measurement invariance across gender on the second-order five-factor model of the German Wechsler intelligence scale for children- fifth edition. *Assessment*. Advance online publication. doi:10.1177/1073191119847762
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. doi:10.1016/j.dr.2016.06.004

- Reynolds, M. R., & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler intelligence scale for children—fifth edition: What does it measure? *Intelligence, 62*, 31-47. doi:[10.1016/j.intell.2017.02.005](https://doi.org/10.1016/j.intell.2017.02.005)
- 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? *The Journal of Genetic Psychology, 176*, 211-234. doi:[10.1080/00221325.2015.1036833](https://doi.org/10.1080/00221325.2015.1036833)
- Rudnev, M., Lytkina, E., Davidov, E., Schmidt, P., & Zick, A. (2018). Testing measurement invariance for a second-order factor: A cross-national test of the alienation scale. *Methods, Data, Analyses, 12*, 47-76. doi:[10.12758/mda.2017.11](https://doi.org/10.12758/mda.2017.11)
- Satorra, A., & Bentler, P. M. (1988). "Scaling corrections for chi-square statistics in covariance structure analysis". In *ASA 1988 proceedings of the business and economic statistics section* (pp. 308-313). Alexandria, VA: American Statistical Association.
- Scheiber, C. (2016). 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? *Journal of Pediatric Neuropsychology, 2*, 79-88. doi:[10.1007/s40817-016-0019-7](https://doi.org/10.1007/s40817-016-0019-7)
- Scheiber, C. (2017). Does the KABC-II display ethnic bias in the prediction of reading, math, and writing in elementary school through high school? *Assessment, 24*, 729-745. doi:[10.1177/1073191115624545](https://doi.org/10.1177/1073191115624545)
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. doi:[10.1177/109442810031002](https://doi.org/10.1177/109442810031002)
- Wechsler, D (2014a). *Wechsler intelligence scale for children* (5th ed.). San Antonio, TX: NCS Pearson.
- Wechsler, D (2014b). *Wechsler intelligence scale for children—fifth edition technical and interpretive manual*. San Antonio, TX: NCS Pearson.