Journal of Psychoeducational Assessment
2014, Vol. 32(7) 597–609

DOI: 10.1177/0734282914538802
jpa.sagepub.com
SAGE

Longitudinal Invariance of the Wechsler Intelligence Scale for Children–Fourth Edition in a Referral Sample

Lindsay P. Richerson[1], Marley W. Watkins[2], and A. Alexander Beaujean[2]

Abstract

Measurement invariance of the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV) was investigated with a group of 352 students eligible for psychoeducational evaluations tested, on average, 2.8 years apart. Configural, metric, and scalar invariance were found. However, the error variance of the Coding subtest was not constant across time, allowing only partial strict invariance. This indicates that the WISC-IV (a) was measuring similar constructs at both test occasions, (b) constructs had the same meaning across time, (c) scores that changed across time can be attributed to change in the constructs being measured and not to changes in the structure of the test itself, and (d) measures the same constructs equally well across time with the possible exception of Processing Speed due to the noninvariance of the Coding subtest's residual variance. This investigation provided support for intelligence as an enduring trait and for the validity of the WISC-IV.

Keywords

WISC-IV, intelligence, invariance, special education, longitudinal

Of all psychological tests, standardized intelligence tests are some of the most widely used by psychologists (Wilson & Reschly, 1996). School psychologists in particular often use standardized intelligence tests as one component of a psychoeducational evaluation for the determination of special education eligibility (Suzuki & Valencia, 1997). Among the available standardized intelligence tests, the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; Wechsler, 2003a) is the most widely used (Strauss, Sherman, & Spreen, 2009). Given that special education eligibility decisions result in relatively long-term placements and may not be beneficial to some children (Morgan, Frisco, Farkas, & Hibel, 2010), strong construct validity evidence is especially important.

[1]Arizona State University, Phoenix, USA
[2]Baylor University, Waco, TX, USA

Corresponding Author:
Marley Watkins, Baylor University, One Bear Place #97301, Waco, TX 76798-7301, USA.
Email: Marley_Watkins@baylor.edu

As intelligence is thought to be an enduring trait (Hunt, 2011), the WISC-IV should evince similar factor structures over time to ensure that the same traits are being measured with equal accuracy across time (Dimitrov, 2010). Unfortunately, there have only been four longitudinal factor analyses of WISC scores during the past 45 years. In the first, the WISC factor structure was investigated with a sample of 153 preschool-age children who were administered the WISC and followed up 1 year later with another administration of the WISC (Osborne, 1965). Using an exploratory factor analysis (EFA) with varimax rotation, the factor structure changed from preschool to first grade. Specifically, there were 8 factors at the first administration and 10 factors at the second administration. However, this study included children who were not of appropriate age for the WISC. In addition, the methodology of this study is problematic as the subtests were split into two, three, or four parts to create additional variables and the EFA methods were suboptimal (Gorsuch, 2003). Because of these limitations, the results of this study should be regarded with caution. Similar techniques and results were reported by Osborne, Anderson, and Bashaw (1967) for the WISC with the same fatal limitations.

In the third study, the WISC-R factor structure was examined using a longitudinal design with a sample of children ($N = 322$) eligible for special education services across a span of approximately 3 years (Juliano, Haddad, & Carroll, 1988). This study enrolled children who were identified as either White or Black; other ethnicities were not included. Results indicated that for students who were administered the Digit Span subtest at test and retest ($n = 229$), a three-factor solution (Verbal, Perceptual, and Freedom from Distractibility) was identified for all groups. Coefficients of congruence were used to quantify similarity between groups, and indicated that the three-factor solution remained stable for children with learning disabilities across the 3-year time span regardless of sex or ethnicity.

The fourth longitudinal factor analysis investigated the factor structure of the WISC-III with 177 students classified as a child with a specific learning disability (SLD), a serious emotional disability (SED), mental retardation (MR), or other disabilities (Watkins & Canivez, 2001). These students were twice administered the WISC-III approximately 3 years apart. Four models were initially evaluated using confirmatory factor analysis (CFA) and the first-order, four-factor model was accepted as the best fitting model for both test and retest occurrences. Test and retest data were also analyzed for invariance of the factor structure across time. Initially, all factor loadings, factor variances, factor covariances, and subtest error variances were constrained to be equal; however, this model had inferior fit in comparison with a baseline model. It was determined that this misfit was likely due to the error variances for three subtests (Vocabulary, Coding, and Arithmetic). Upon releasing those constraints, the model fit was significantly improved. These results indicated that the WISC-III measured the same constructs across time and that the constructs were manifested in the same way across groups.

There have been no investigations of the longitudinal factorial invariance of the WISC-IV. Cross-sectional analysis of the WISC-IV has supported the assumption of longitudinal invariance (Keith, Fine, Taub, Reynolds, & Kranzler, 2006), but cross-sectional analyses may not be adequate for detecting change over time (Willett, Singer, & Martin, 1998). Thus, there is no evidence regarding the factorial invariance of the WISC-IV across time for the *same* individuals. If longitudinal factorial invariance exists, differences in obtained WISC-IV test–retest scores can be unequivocally attributed to respondents changing on the underlying constructs being measured. In the absence of longitudinal factorial invariance, WISC-IV-obtained test–retest scores cannot be compared because changes in test scores could be due to a myriad of reasons other than changes in the respondents' standing on the underlying constructs (Dimitrov, 2010). In that situation, the use of WISC-IV scores for identification of children with disabilities would be suspect. Therefore, the current study will use CFA techniques to examine the temporal stability of the factor structure of the WISC-IV in a clinically referred sample.

Method

Participants

Three hundred fifty-two students (66% males) who were twice administered the WISC-IV, with all 10 core subtests administered at each test session, served as participants in the current study. Participant ages ranged from 6.1 to 14.11 years at first testing and 7.5 to 16.6 years at second testing with an average test–retest interval of 2.84 years. Reported ethnic breakdown of the sample was 79% White, 11% Hispanic, 6% Black, and 4% Other. Special education placement was determined by local multidisciplinary evaluation teams following state regulations. Special education diagnosis on initial evaluation included 66% SLD, 9% other health impairment (OHI; attention-deficit/hyperactivity disorder [ADHD]), 8% SED, 5% nonhandicapped, 4% autism, 2% MR, 3% OHI (non-ADHD), and 3% other. To preserve respondents' privacy, no other information was collected.

Instrument

The WISC-IV is an individually administered intelligence test for children between the ages of 6 and 16 years. The WISC-IV consists of 15 subtests, 10 core and 5 supplemental, each with a mean of 10 and a standard deviation of 3. The 10 core subtests are used to form a Full Scale Intelligence Quotient (FSIQ) score as well as four index scores: Verbal Comprehension Index (VCI; Similarities, Vocabulary, and Comprehension), Perceptual Reasoning Index (PRI; Block Design, Matrix Reasoning, and Picture Concepts), Working Memory Index (WMI; Digit Span and Letter-Number Sequencing), and Processing Speed Index (PSI; Coding and Symbol Search). The FSIQ and index scores have a mean of 100 and a standard deviation of 15.

There has been some debate about the factor structure of the WISC-IV. The technical manual reported that a first-order, four-factor oblique structure fit the core subtests the best (Wechsler, 2003b), mapping onto the VCI, WMI, PSI, and PRI index scores. Others studies have found that a higher order (Keith et al., 2006) or bifactor (Watkins, 2006) general intelligence factor (*g*) should also be considered, as it explained more of the subtest covariance than any first-order factor (Bodin, Pardini, Burns, & Stevens, 2009; Watkins, 2006, 2010; Watkins, Wilson, Kotz, Carbone, & Babula, 2006).

Procedure

Following Institutional Review Board (IRB) and school district approval, special education files in two participating Southwestern school districts were reviewed and relevant WISC-IV scores were extracted. In total, there were 457 students who were twice administered the WISC-IV. However, only 352 students had complete subtest scores at both test and retest. School district demographics were collected from information provided by the National Center for Educational Statistics (2012). The first district comprised approximately 84% non-Hispanic or Latino students, with 6% of their students identified as English Language Learners. The second district comprised approximately 88% non-Hispanic or Latino students, with 4% of their students' identified as English Language learners.

Analyses

Model specification. CFA will allow a robust examination of the invariance of the WISC-IV structure across time (Byrne & Stewart, 2006; Millsap & Cham, 2012). When examining factorial invariance, the first step is to determine the baseline factor structure within each testing occasion

Table 1. Levels of Measurement Invariance.

Model	Title	Description	Comparisons allowed
1	Configural	The factor model for all groups is the same. No parameter constraints are imposed.	None
2	Weak/metric	1 + all factor pattern coefficients are constrained to be the same between groups (but can vary within a group)	Factor (co)variances (weak evidence)
3	Strong/scalar	2 + all intercepts are constrained to be the same between groups (but can vary within a group)	Factor means, factor (co)variances (strong evidence)
4		3 + constrain the factor variances/covariances to be the same across groups	Reliability (necessary, but not sufficient)
5	Strict	3 + measurement error variances/covariances are constrained to be the same between groups (but can vary within a group)	Reliability (sufficient in conjunction with Model 4)

(van de Schoot, Lugtig, & Hox, 2012). For this study, we tested three models: (a) four oblique first-order factors representing the VCI, PRI, WMI, and PSI; (b) a higher order factor model with one second-order factor and four first-order factors; and (c) a bifactor model with one general factor and four orthogonal domain-specific factors. For bifactor model identification, we constrained the loadings for the WMI subtests to be equal and the loadings for the PSI subtests to be equal.

Testing invariance. Testing invariance across time is similar to testing invariance across groups, except the covariances between like indicator variables' uniquenesses and common factors across measurement occasions are sometimes included in the model due to domain-specific covariance not accounted for by the factor model (McArdle, 2009). Consequently, Vandenberg and Lance (2000) noted that there are two ways to assess measurement invariance with longitudinal data. The first is to treat the data at the different occasions as if they came from two separate groups and conduct invariance assessment as a typical multigroup model. Although this model is the more parsimonious of the two, it cannot account for correlated residuals or factors across time.

The second approach is to treat the data as if they come from a single sample, similar to traditional repeated measures ANOVAs. This way of assessing invariance posits as many factor models as there are time points, and allows across-occasion covariances for each indicator's residual variance and each common factor. A disadvantage of this approach is that the input covariance matrix is made up of both the within-occasion and between-occasion covariances, which sometimes results in poor model fit and improper solutions. However, the same levels of invariance are investigated for either the single-sample or multiple-group approach (see Table 1).

Determining model fit. Researchers (e.g., Byrne & Stewart, 2006) have suggested two sets of criteria for testing factorial invariance. The traditional perspective examines the change in χ^2 ($\Delta\chi^2$) across nested models. If, as the models grow more restrictive, the χ^2 values do not significantly change (using a given α level), this is evidence that the more restrictive model fits the data as well as the less restrictive model; thus, the more restrictive (i.e., more parsimonious) model should be favored over the less restrictive one.

The use of χ^2 values has been criticized because of their sensitivity to sample size (Byrne & Stewart, 2006). Cheung and Rensvold (2002) and Meade, Johnson, and Braddy (2008) argued that some alternative fit indices (AFIs) are not as susceptible to this problem. Specifically, they found that the comparative fit index (CFI) and McDonald's (1989) noncentrality index (Mc)

Table 2. Descriptive Statistics for Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV) Scores of 352 Students Twice-Tested for Special Education Eligibility.

	M		*SD*	
Variable	Test	Retest	Test	Retest
Block design	9.19	8.67	2.80	2.98
Similarities	8.77	9.18	2.62	2.80
Digit span	7.98	7.82	2.61	2.58
Picture concepts	9.53	9.99	3.31	3.00
Coding	8.41	7.52	3.15	2.90
Vocabulary	8.60	8.44	2.65	2.74
Letter-number sequencing	8.05	8.17	2.81	3.12
Matrix reasoning	9.08	9.10	2.95	3.09
Comprehension	8.85	8.92	2.71	2.60
Symbol search	8.44	8.67	3.24	3.10
Verbal comprehension	92.5	93.0	12.7	13.2
Perceptual reasoning	95.5	95.4	15.0	15.7
Working memory	88.3	88.0	13.0	14.2
Processing speed	91.3	89.3	15.2	15.0
Full scale IQ/*g*	90.3	89.9	13.6	14.5

Note.IQ = intelligence quotient.

were more robust. Thus, the second set of evaluations criteria takes a practical perspective and recommends that invariance be based on two criteria: (a) The multigroup factor model exhibits an adequate fit to the data, and (b) the change in values for AFIs (e.g., ΔCFI, ΔMc) is negligible.

Based on Byrne and Stewart's (2006) recommendations, this study used two sets of fit indices: one to assess overall model fit and the other to assess change in model fit between two models. As Hu and Bentler (1999) suggested, we used multiple fit indices for both. For this study's criteria of overall model-data fit, we used the following: (a) root mean square error of approximation (RMSEA) ≤ .08; (b) standardized root mean square residual (SRMR) ≤ .08, and (c) CFI ≥ .96 (Hu & Bentler, 1999; Yu, 2002). To test the change in fit between nested models, we used the ΔCFI and ΔMc (Meade et al., 2008). Cheung and Rensvold (2002) suggested .01 as the threshold for ΔCFI and .02 as the threshold for ΔMc.

For both overall model fit as well as change in model fit, we looked for patterns in the fit statistics and judged acceptance/rejection of the specific model based on the majority of the indices. All analyses were done in R (R Development Core Team, 2012) using the *lavaan* (Rosseel, 2012) and *psych* (Revelle, 2012) statistical packages.

Results

Data Inspection

Descriptive statistics for WISC-IV subtest, factor, and IQ scores at test and retest for this referred sample are reported in Table 2 and correlations between subtests at test and retest are provided in Table 3. These results indicate that the current sample exhibited slightly lower and more variable scores than the normative sample of the WISC-IV (Wechsler, 2003b). Similar score patterns have been observed in other clinical samples (Watkins et al., 2006). The univariate score distributions from the current sample appear to be relatively normal across both test administrations (West,

Table 3. Correlations of WISC-IV Subtests at Test and Retest.

	VC	SI	CO	BD	PCn	MR	DS	LN	CD	SS
VC	.69	.63	.64	.38	.43	.39	.40	.44	.18	.33
SI	.70	.59	.48	.28	.35	.34	.30	.32	.06	.34
CO	.60	.53	.50	.23	.40	.30	.24	.36	.19	.31
BD	.37	.45	.39	.71	.43	.53	.46	.41	.24	.37
PCn	.40	.45	.40	.46	.47	.46	.36	.38	.23	.38
MR	.44	.51	.41	.65	.56	.62	.39	.45	.14	.37
DS	.39	.39	.40	.45	.30	.47	.61	.47	.20	.39
LN	.48	.41	.48	.46	.36	.47	.53	.49	.30	.37
CD	.15	.11	.37	.30	.28	.29	.29	.32	.52	.44
SS	.27	.28	.37	.40	.33	.42	.32	.41	.62	.54

Note. Test correlations are in the upper triangle, Retest correlations are in the lower triangle, and test–retest correlations are on the diagonal. WISC-IV = Wechsler Intelligence Scale for Children–Fourth Edition; VC = Vocabulary; SI = Similarities; CO = Comprehension; BD = Block Design; PCn = Picture Concepts; MR = Matrix Reasoning; DS = Digit Span; LN = Letter-Number Sequencing; CD = Coding; SS = Symbol Search.

Table 4. Fit Statistics for Alternative Baseline Models at Test and Retest.

Model	χ^2	*df*	CFI	RMSEA	SRMR
Test					
Four oblique factors	61.53	29	.96	.06	.03
Second-order	62.25	31	.97	.05	.03
Bifactor	58.91	27	.96	.06	.03
Retest					
Four oblique factors	90.94	29	.94	.08	.04
Second-order	98.09	31	.94	.08	.05
Bifactor	77.37	27	.95	.07	.04

Note. All statistics based on scaled χ^2 statistic and significant using $\alpha = .01$. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's noncentrality index.

Finch, & Curran, 1995). In addition, examination of each variable's associated histogram indicated that the sample appears to generally follow the shape of a normal distribution. Nevertheless, we used maximum likelihood parameter estimators with standard errors and a mean-adjusted chi-square test statistic that are robust to nonnormality (Satorra & Bentler, 2001).

Factor Models

Table 4 contains the fit statistics for the three alternative models within each testing occasion. Not unexpectedly, the models fit relatively similarly at both time points (Murray & Johnson, 2013). Chen, West, and Sousa (2006) suggested that when examining invariance, the bifactor model is better than a second-order factor model because the bifactor model allows for tests of invariance of the domain-specific factors as well as the general factor. In contrast, a second-order model only allows for direct tests of invariance for the second-order factor, as the first-order factors are represented by disturbances. Consequently, we chose the bifactor model to use for the invariance assessment. Figure 1 displays the bifactor model. This choice was corroborated by an EFA as per Carroll (1993), which produced similar orthogonal structures.

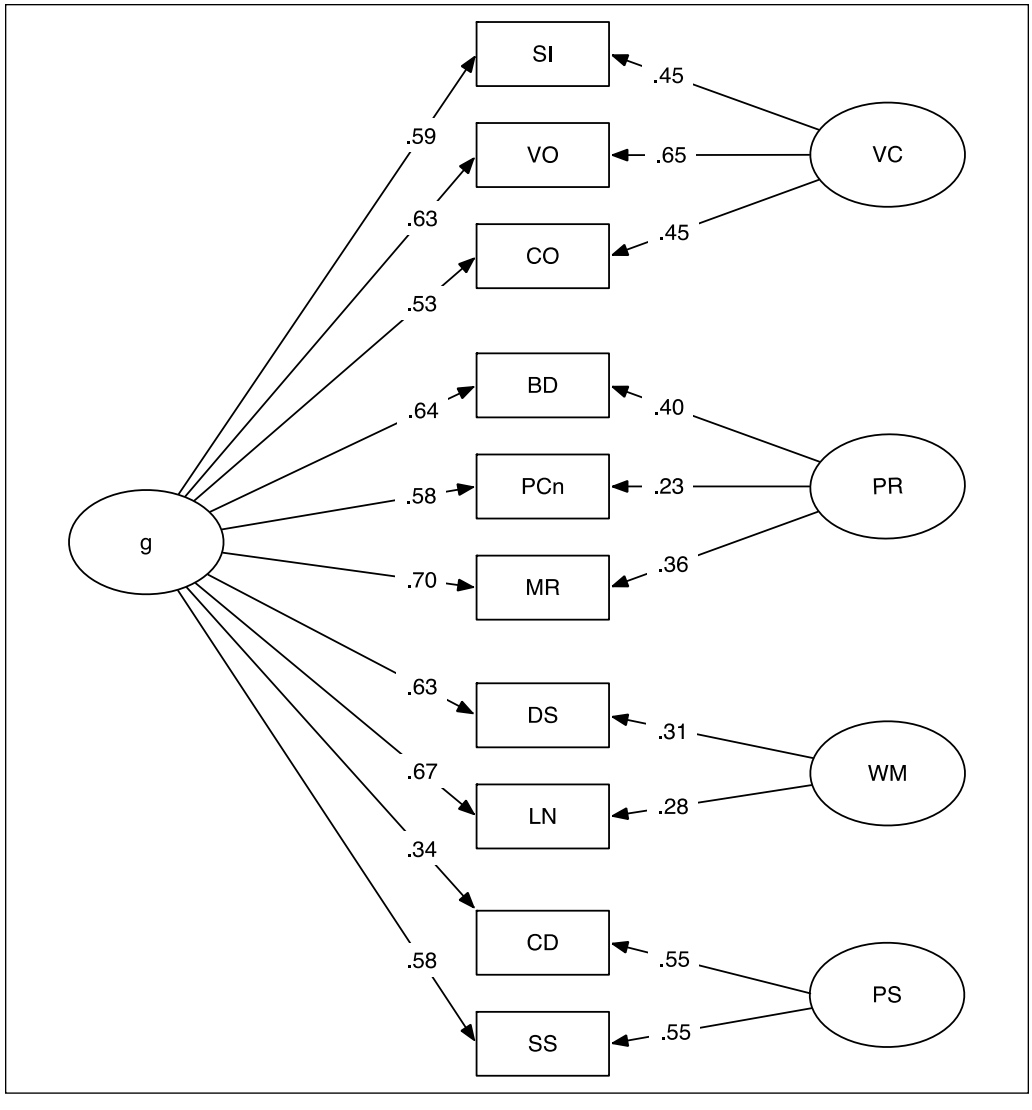

Figure 1. Bifactor model with orthogonal domain-specific group factors.
Note. All values are standardized and from Model 5a. Residual variances are not shown. SI = Similarities; VO = Vocabulary; CO = Comprehension; BD = Block Design; PCn = Picture Concepts; MR = Matrix Reasoning; DS = Digit Span; LN = Letter-Number Sequencing; CD = Coding; and SS = Symbol Search; VC = Verbal Comprehension factor; PR = Perceptual Reasoning factor; WM = Working Memory factor; PS = Processing Speed factor.

Invariance

The first step in testing the measurement invariance hierarchy was to assess configural invariance using both the single-sample and multiple-group approach (van de Schoot et al., 2012). Initially, we tested for configural invariance using the multiple-group approach, which does not allow residual or factor variances to covary across time (see Model 1a in Table 5). Although the χ^2 value was statistically different than zero, the AFIs indicated that the model fit the data relatively well. Subsequently, we examined configural invariance using the single-sample approach, allowing the common factors and residual variances from the same indicators to covary across the two

Table 5. Fit Statistics for Invariance Models.

Model	χ^2	*df*	CFI	RMSEA	SRMR	Mc
1a—Configural (multigroup)	136.21	54	.957	.07	.04	.943
1b—Configural (single sample)	219.95	139	.969	.04	.04	.891
2—Metric	248.14	152	.963	.04	.05	.872
3—Scalar	259.86	157	.960	.04	.05	.864
4—Latent variances	272.35	162	.957	.04	.06	.855
5—Strict invariance	300.29	172	.950	.05	.05	.833
5a—Strict invariance (partial)	286.91	171	.955	.04	.06	.848

Note. All statistics based on scaled χ^2 statistic. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's noncentrality index.

Table 6. Change in Fit Statistics for Invariance Models.

Comparison	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔMc
Model 1b vs. 2	28.19	13	.01	.006	.019
Model 2 vs. 3	11.72	5	.04	.003	.008
Model 3 vs. 4	12.49	5	.03	.003	.009
Model 4 vs. 5	27.93	10	.01	.007	.022
Model 4 vs. 5a	14.55	9	.10	.002	.007

Note. See Table 1 for invariance model descriptions. $\Delta\chi^2$ based on scaled difference (Satorra & Bentler, 2001). CFI = comparative fit index; Mc = McDonald's noncentrality index.

time points (Model 1b). For all fit indices except Mc, the single-sample model showed a better fit to the data than the multiple-group model. Consequently, we used the single-sample model as our baseline for subsequent tests of invariance.

We next examined metric/weak invariance (van de Schoot et al., 2012), which constrains the factor loadings between groups (Model 2). This analysis allows factor variances between groups to vary, so we constrained the following loadings for identification: (a) Vocabulary and Coding were constrained to one for the domain-specific factors; (b) Similarities was constrained to one for the general factor; and (c) both loadings for the WM factor and the PS factor were constrained to one because each factor comprised only two subtests.[1] The values for the CFI, RMSEA, and SRMR indices indicated that this model fit the data relatively well. Moreover, the $\Delta\chi^2$, ΔCFI, and ΔMc values indicated that the model did not fit worse than Model 1b using the Cheung and Rensvold (2002) criteria (see Table 6). Substantiation of metric invariance was also obtained from the EFA (Horn & McArdle, 1992; Lorenzo-Seva & ten Berge, 2006), with congruence coefficients that ranged from good (.97) to excellent (.99) according to the guidelines provided by MacCallum, Widaman, Zhang, and Hong (1999).

A number of measurement researchers agree that achieving both configural and metric factorial invariance is enough evidence to determine that a measure is invariant across time (Bentler, 2005; Widaman & Reise, 1997) and that further invariance testing is discretionary (Vandenberg & Lance, 2000; Wu, Li, & Zumbo, 2007) or unwarranted (Selig, Card, & Little, 2008). Others believe that strict invariance is required, especially when tests are used for individual decisions (Meredith & Teresi, 2006). Accordingly, this study continued to evaluate measurement invariance by addressing both strong/scalar and strict levels of invariance.

We next examined scalar/strong invariance (van de Schoot et al., 2012), which constrains the manifest variables' intercepts between groups but allows the latent variables' means to differ

between groups (Model 3). All the AFIs indicated that the model fit the data relatively well. Moreover, the $\Delta\chi^2$, ΔCFI, and ΔMc values all indicated that the model fit no worse than the metric invariance model.

Next, we tested the latent variables' variances across test and retest (van de Schoot et al., 2012). We constrained all the latent variable's variances to be one and allowed the loadings for WMI and PSI factors to be a value different than one. All the model fit indices indicated that the model fit the data relatively well. The $\Delta\chi^2$, ΔCFI, and ΔMc values all indicated that the model fit no worse than the scalar invariance model.

The strict invariance model, which constrains the residual variances across groups (van de Schoot et al., 2012), was tested next (Model 5). The SRMR and RMSEA indicated that the model fit the data relatively well, but the $\Delta\chi^2$, ΔCFI, and ΔMc values indicated that that the model fit worse than the previous model. Thus, the model does not appear to have complete strict invariance across time. An examination of the residual variances found the Coding subtest to be most disparate. We removed the equality constraints for the Coding subtest and refit the model (Model 5a). All the model fit indices indicated that the revised model fit the data relatively well. The $\Delta\chi^2$, ΔCFI, and ΔMc values all indicated that the model fit no worse than the prior test of latent variances. Thus, the model exhibited partial strict invariance, indicating that any differences between the means and variances of the WISC-IV subtests was due solely to differences in the constructs that they measure. Thus, with the exception of the Coding subtest, "all group differences on the measured variables are captured by, and attributable to, group differences on the common factors" (Widaman & Reise, 1997, p. 296). The final model is illustrated in Figure 1.

Discussion

The goal of the current study was to investigate measurement invariance of the WISC-IV for a group of 352 students eligible for psychoeducational evaluations tested, on average, 2.8 years apart. Using CFA methods, the bifactor model exhibited partial strict invariance across time, with the error variance of the Coding subtest being the only residual variance that differed across time.

Verification of configural invariance indicates that the same factor structure was maintained across time. Thus, there was the same number of latent variables, indicator variables, and pattern of fixed and estimated parameters at both test and retest. This indicates that the WISC-IV was measuring similar constructs at both test and retest occasions. Configural invariance is considered to be the least restrictive test of similarity of factors across time (Dimitrov, 2010).

The achievement of metric invariance means that corresponding factor loadings (i.e., pattern coefficients) were equivalent across time. That is, each subtest loaded equivalently on its respective factors at both test and retest occasions. Thus, the constructs being measured were equivalent at both test and retest. This provides evidence that the observed WISC-IV scores (e.g., FSIQ, VCI, PRI, etc.) were assessing factors of the WISC-IV (e.g., *g*, VC, PR, etc.) in the same way at both test and retest (Horn & McArdle, 1992; Wu et al., 2007).

Attaining scalar/strong invariance indicates that factor means and variances can be compared across time (Dimitrov, 2010). Therefore, any change in observed WISC-IV test scores (e.g., FSIQ, VCI, PRI, etc.) across time can be attributed to change in the constructs being measured (e.g., *g*, VC, PR, etc.) and not to changes in the structure of the test itself. Thus, students with the same ability at either test occasion achieved the same manifest scores on the WISC-IV, allowing valid comparisons of mean scores and correlations across groups (Horn & McArdle, 1992).

A model with partial strict invariance indicates that the latent variables the WISC-IV is measuring, with the possible exception of Processing Speed due to the noninvariance of the Coding subtest's residual variance, were measured with equal precision at both test occasions. The error variance of the WISC-III Coding subtest was also found to lack longitudinal invariance (Watkins & Canivez, 2001). Thus, differences in WISC-IV obtained scores across time (with the exception

of Coding) were due to differences in their latent means (Dimitrov, 2010). This supports the hypothesis that the WISC-IV measures the same constructs equally well across time.

Limitations

As with all research, there are a number of limitations in the current study. The greatest of these limitations is the sample. Although a sample of 352 students is typically considered to be large, this is a relatively small sample for factorial invariance testing of complex structures. Ideally, a larger sample is desired when completing these types of analysis (Byrne, 2012). In addition, the sample used in this study was from two school districts in one Southwestern state and thus may not be generalizable to other regions. Furthermore, the sample consisted solely of students twice referred for a psychoeducational evaluation for special education eligibility. The characteristics that resulted in two WISC-IV administrations may have been unique. A final limitation of this study is the method of data collection. As the data was collected from archived special education records, administration and recording accuracy of the individual psychologists who administered the WISC-IV had to be assumed.

Conclusion

Although the longitudinal structural stability of the WISC-IV has not previously been investigated, cross-sectional measurement has found it to be consistent across ages 6 to 16 years (Keith et al., 2006). Likewise, the temporal stability of other cognitive ability test scores has been demonstrated with children (Watkins & Canivez, 2001) as well as adults (Reeve & Bonaccio, 2011). The current study demonstrated that changes in WISC-IV scores across time can be attributed to change in the constructs being measured and not to change in the structure of the test itself. These results provide support for intelligence as an enduring trait (Hunt, 2011) and for the validity of the WISC-IV. However, obtained factor index scores are not pure measures of their underlying constructs because each obtained index score is influenced by *g* as well as error. For example, about 60% of the variance in the VCI score is due to *g* (Schneider, 2013). This complex relationship between latent and obtained scores should be considered when interpreting WISC-IV subtest, index, and full scale scores (DeMars, 2013).

Acknowledgment

The contributions of Dr. John Balles and Dr. Christa Lynch are gratefully acknowledged.

Authors' Note

This study is based on the dissertation of the first author. Dr. Richerson is now with the Scottsdale Unified School District, Scottsdale, Arizona.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. We also fit a model constraining the variance of Working Memory (WM) and Processing Speed (PS) to one and estimating the factor loadings of the WM and PS factors. The difference in fit

was minimal: $\chi^2 = 251.66$, $df = 154$, comparative fit index (CFI) = .962, root mean square error of approximation (RMSEA) = .04, standardized root mean square residual (SRMR) = .05, and McDonald's noncentrality index (Mc) = .870.

References

Bentler, P. M. (2005). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Bodin, D., Pardini, D. A., Burns, T. G., & Stevens, A. B. (2009). Higher order factor structure of the WISC-IV in a clinical neuropsychological sample. *Child Neuropsychology*, *15*, 417-424. doi:10.1080/09297040802603661

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications and programming*. New York, NY: Routledge.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, *13*, 287-321. doi:10.1207/s15328007sem1302_7

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

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*, 189-225. doi:10.1207/s15327906mbr4102_5

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. doi:10.1207/S15328007SEM0902_5

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*, 354-378. doi:10.1080/15305058.2013.799067

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121-149.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2., pp. 143-164). Hoboken, NJ: Wiley.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *11*, 117-144.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi:10.1080/10705519909540118

Hunt, E. (2011). *Human intelligence*. New York, NY: Cambridge University Press.

Juliano, J. M., Haddad, F. A., & Carroll, J. L. (1988). Black and white, female and male children classified as learning-disabled. *Journal of School Psychology*, *26*, 317-325.

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

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57-64. doi:10.1027/1614-1881.2.2.57

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84-99.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577-605. doi:10.1146/annurev.psych.60.110707.163612

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*, 97-103. doi:10.1007/bf01908590

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568-592. doi:10.1037/0021-9010.93.3.568

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, S69-S77.

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109-126). New York, NY: Guilford Press.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *Journal of Special Education*, *43*, 236-254.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*, 407-422. doi:10.1016/j.intell.2013.06.004

National Center for Educational Statistics. (2012, February 3). *Elementary/secondary information system*. Available from http://nces.ed.gov

Osborne, R. T. (1965). Factor structure of the Wechsler Intelligence Scale for Children at preschool level and after first grade: A longitudinal study. *Psychological Reports*, *16*, 637-644.

Osborne, R. T., Anderson, H. E., & Bashaw, W. L. (1967). The stability of the WISC factor structure at three age levels. *Multivariate Behavioral Research*, *2*, 443-451.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Reeve, C. L., & Bonaccio, S. (2011). On the myth and the reality of the temporal validity degradation of general mental ability test scores. *Intelligence*, *39*, 255-272.

Revelle, W. (2012). psych: Procedures for psychological, psychometric, and personality research (Version 1.2.4) [computer software]. Evanston, IL: Northwestern University.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1-36.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514. doi:10.1007/bf02296192

Schneider, W. J. (2013). What if we took our models seriously? Estimating latent scores in individuals. *Journal of Psychoeducational Assessment*, *31*, 186-201. doi:10.1177/0734282913478046

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modeling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. van de Vijver, D. A. van Hemert, & Y. H. Poortinga (Eds.), *Analysis of individuals and cultures* (pp. 93-119). New York, NY: Lawrence Erlbaum.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2009). *A compendium of neuropsychological tests: Administration, norms, and commentary*. New York, NY: Oxford University Press.

Suzuki, L. A., & Valencia, R. R. (1997). Race-ethnicity and measured intelligence: Educational implications. *American Psychologist*, *52*, 1103-1114.

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

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-492. doi:10.1080/17405629.2012.686740

Watkins, M. W. (2006). Orthogonal higher order structure of the Wechsler Intelligence Scale for Children–Fourth Edition. *Psychological Assessment*, *18*, 123-125. doi:10.1037/1040-3590.18.1.123

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children–Fourth Edition among a national sample of referred students. *Psychological Assessment*, *22*, 782-787. doi:10.1037/a0020043

Watkins, M. W., & Canivez, G. L. (2001). Longitudinal factor structure of the WISC-III among students with disabilities. *Psychology in the Schools*, *38*, 291-298.

Watkins, M. W., Wilson, S. M., Kotz, K. M., Carbone, M. C., & Babula, T. (2006). Factor structure of the Wechsler Intelligence Scales–Fourth Edition among referred students. *Educational and Psychological Measurement*, *66*, 975-983. doi:10.1177/0013164406288168

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children–Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003b). *Wechsler Intelligence Scale for Children–Fourth Edition technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. A. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56-75). Newbury Park, CA: SAGE.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.),

The science of prevention: Methodological advances from alcohol and substance abuse research (pp. 281-322). Washington, DC: American Psychological Association.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395-426.

Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review*, *25*, 9-23.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*, 1-26.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles.