Personality and Individual Differences 86 (2015) 1–5

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Factor invariance between genders on the Wechsler Intelligence Scale for Children–Fifth Edition

Hsinyi Chen [a,*], Ou Zhang [b], Susan Engi Raiford [b], Jianjun Zhu [b], Lawrence G. Weiss [b]

[a] Department of Special Education, National Taiwan Normal University, Taipei, Taiwan, ROC
[b] Clinical Assessment, Pearson, San Antonio, TX, USA

ARTICLE INFO

Article history:
Received 26 March 2015
Received in revised form 8 May 2015
Accepted 13 May 2015
Available online xxxx

Keywords:
Factorial invariance
Gender
Wechsler scales

ABSTRACT

This study investigated the factorial invariance of the Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V) between samples of male and female children. A higher-order 5-factor model was tested on a nationally-representative sample of 2200 children aged 6 to 16 years. The results demonstrated full factorial invariance between genders. The WISC-V subtests demonstrate the same underlying theoretical latent constructs, the same strength of relationships among factors and subtests, the same validity of each first-order factor, and the same communalities, regardless of the gender, thus supporting the same interpretive approach and meaningful comparisons of the WISC-V between male and female children.

1. Introduction

Wechsler tests are among the most widely used intelligence instruments worldwide (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Bowden, 2013; Rabin, Barr, & Burton, 2005). Roughly twenty countries have adapted and standardized Wechsler intelligence scales to date (Camara, Nathan, & Puente, 2000; Georgas, Weiss, van de Vijver, & Saklofske, 2003). The Wechsler intelligence scales are revered because of their psychometric properties and practical relevance (Groth-Marnat, 2009, p. 119).

Invariance is a fundamental property of any instrument that may be used to compare individuals from subpopulations. Meaningful comparisons can be made only if the measures are comparable and a lack of evidence for measurement invariance hinders the ability of the measure to be used in comparisons among groups (AERA, APA, NCME, 2014; Chen, Sousa, & West, 2005; Drasgow, 1984, 1987; Horn & McArdle, 1992; Millsap & Kwok, 2004; Rock, Werts, & Flaugher, 1978; Vandenberg & Lance, 2000). The Wechsler intelligence scales are frequently utilized in the course of psychoeducational assessments (Flanagan & Kaufman, 2004; Prifitera, Saklofske, & Weiss, 2005, 2008; Sattler & Dumont, 2004; Weiss, Saklofske, Prifitera, & Holdnack, 2008). Implicit in such common practice is the assumption that Wechsler intelligence scale scores have the same meaning for children in various subpopulations. Thus, investigating the measurement invariance of Wechsler intelligence scales is crucial.

The Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V; Wechsler, 2014a) is the latest edition of Wechsler's test of child intelligence, which has its roots in the Wechsler Bellevue Form II published in 1946 by Wechsler. The WISC-V is a major revision of the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003), and it does incorporate many significant changes. Chief among these is that compared to the four-factor model utilized in the WISC-IV, the WISC-V utilizes a new five-factor scoring framework, with the factors as follows: Verbal Comprehension (VCI), Visual Spatial (VSI), Fluid Reasoning (FRI), Working Memory (WMI), and Processing Speed (PSI) (Wechsler, 2014a). For the past decade, studies worldwide have shown firm support for WISC-IV measurement invariance between genders (Chen & Zhu, 2008), and across various cultures (Chen, Keith, Weiss, Zhu, & Li, 2010), ages (Keith, Fine, Taub, Reynolds, & Kranzler, 2006), and clinical status (Chen, Hung, Chen, Zhu, & Keith, in press; Chen & Zhu, 2012; Weiss, Keith, Zhu, & Chen, 2013). In addition, studies of the WISC-IV found support for a five-factor structure among the normative (Keith et al., 2006; Weiss et al., 2013) and clinical samples (Weiss et al., 2013), and the *WISC-V Technical and Interpretive Manual* (Wechsler, 2014b) provided evidence supporting this new structure in the new version, but questions about consistency of measurement across subpopulations remain to be answered for the WISC-V (Canivez & Watkins, in press).

Among all possible subgroup classifications, gender invariance is recognized as fundamental for measurements in various domains (Atienza, Balaguer, & Garcia-Merita, 2003; Byrne, Baron, & Campbell, 1993; Cheng & Watkins, 2000; Richardson, Huan, Ege, Suh, & Rice, 2014; Rusticus & Hubley, 2006). For data from males and females are usually combined when substantive applied studies of the Wechsler

* Corresponding author. Tel.: +886 2 77345011; fax: +886 2 23413061.
E-mail address: hsinyi@ntnu.edu.tw (H. Chen).

http://dx.doi.org/10.1016/j.paid.2015.05.020
0191-8869/

intelligence scales are conducted empirically, gender invariance certainly is an essential issue pertaining to WISC-V. Besides, we need evidence showing that the WISC-V is not a biased tool against gender and thus any future gender difference based on this instrument could be genuine.

This study investigates gender invariance with large samples with considerable variation. Specially, we evaluated whether the WISC-V subtests measure latent abilities in the same manner for both male and female children.

2. Method

2.1. Participants

We analyzed the WISC-V standardization responses from 2200 children (males N = 1009; females N = 1101). This nationally representative sample was divided into 11 age groups from ages 6 to 16, with 200 children in each age group. This sample was carefully selected to match the 2012 United States Census on geographic region, gender, parent education level, and race/ethnicity. A detailed description of this sample is provided in the WISC-V manual (Wechsler, 2014b).

2.2. Instrumentation

The WISC-V has 10 primary subtests and six secondary subtests. The 10 primary subtests are Similarities (SI), Vocabulary (VC), Block Design (BD), Visual Puzzles (VP), Matrix Reasoning (MR), Figure Weights (FW), Digit Span (DS), Picture Span (PS), Coding (CD), and Symbol Search (SS). The six secondary subtests are Information (IN), Comprehension (CO), Picture Concepts (PC), Arithmetic (AR), Letter–Number Sequencing (LN), and Cancellation (CA). All composites and subtests have demonstrated good reliability, with average internal consistency reliability estimates ranging from 0.88 to 0.96 for composites, 0.81 to 0.94 for primary subtests, and .82 to .90 for secondary subtests (Wechsler, 2014b, pp.57). We employed all 10 primary subtests and six secondary subtests in this study to ensure adequate markers for reliable latent abilities.

2.3. Analysis of the data

Tests to measure invariance between genders were based on the analysis of covariance structure models using LISREL 8.8 (Jöreskog & Sörbom, 2006). We first checked the normality of each subtest. In both male and female groups, skewness ranged from −.14 to .12, and kurtosis ranged from −.22 to .50. Maximum likelihood estimation is known for robustness (Hu & Bentler, 1998), and is considered adequate for data with a skewness of less than 2 and a kurtosis of less than 7 (West, Finch, & Curran, 1995). Thus, we used maximum likelihood estimation for model estimation.

Prior to invariance analysis, we separately tested the corresponding five-factor baseline model for males and females. The five-factor structure reported in the *WISC-V Technical and Interpretive Manual* (Wechsler, 2014b, p. 83) was used as the hypothesized baseline model. For the 16-subtest version, the baseline model specified a higher-order *g* and five first-order factors. There are four Verbal Comprehension subtests (SI, VC, IN, CO) on the first factor, two Visual Spatial subtests (BD, VP) on the second factor, four Fluid Reasoning subtests (MR, FW, PC, AR) on the third factor, three Working Memory subtests (DS, PS, LN) on the fourth factor, and three Processing Speed subtests (CD, SS, CA) on the fifth factor. The Arithmetic subtest was allowed to be cross-loaded on the Fluid Reasoning, Working Memory, and Verbal Comprehension factors. This five-factor structure is displayed in Fig. 1.

We examined the factorial invariance by testing six levels of nested models to investigate the degree of invariance (Keith, 2015; Meredith, 1993; Vandenberg, 2002; Wicherts & Dolan, 2010). Each level had more constraints than those of the previous level. The initial and weakest level was configural invariance, which assumed the same number of factors and the same overall factor pattern across groups. The second level was first-order factor-loading invariance (or metric/weak factorial invariance). Loadings of subtests on factors were constrained so that factor loadings were equal across groups. When the factor loadings are equal, the scales of the latent variables are the same for both groups, and the unit of measurement is identical. The third level was intercept invariance (or scalar/strong factorial invariance). At this level, any group difference in subtest means result from the true mean differences in latent factors. The subtests have the same intercepts across groups if they have the same latent factor means. The fourth level tested residual invariance (or strict factorial invariance) to examine whether "all group differences on the measured variables are captured by, and attributable to, group differences on the common factors" (Widaman & Reise, 1997, p. 296). These residuals are a combination of subtest-specific unique variance and measurement errors. The fifth level was second-order factor-loading invariance. We assumed that first-order latent factors show the same amount of change in each group for the same increase in *g*. Finally, we tested the invariance of disturbances (factor unique variances) of the first-order factors. Although disturbance invariance is not fundamentally crucial for measurement invariance, it provides substantial information regarding human cognitive abilities across groups. We did not constrain first-order factor intercepts to be equal across groups, because such constraints addressed measurement questions that do not pertain to the current study. For all analyses, we identified the scale of latent factors by fixing a factor loading of each factor to one.

Multiple indices of the model fit were used to evaluate and compare the models (Bentler & Bonett, 1980; Hoyle & Panter, 1995; Hu & Bentler, 1998, 1999; Kline, 2010; Marsh, Balla, & McDonald, 1988; McDonald & Ho, 2002). Single models were jointly evaluated by using the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). An RMSEA value less than .05 corresponded to a good fit, and .08 was considered to be acceptable. SRMR values less than .08 were considered to be good. A value of .95 served as the cutoff point for an acceptable fit of all indices ranging from 0 to 1, with 1 indicating a perfect fit. Change in the chi-square ($\Delta\chi^2$) value was used to evaluate competing nested models (Bentler & Bonett, 1980). The Akaike information criterion (AIC) and sample size adjusted Bayesian Information Criterion (aBIC) were used for comparisons of competing nested and non-nested models (Kaplan, 2000; Loehlin, 2004), with lower values indicating a superior fit. The aBIC has a more substantial reward for parsimony compared with the AIC.

To determine evidence of invariance, consensus is scant regarding the most appropriate criterion (Byrne & Stewart, 2006; Meade, Johnson, & Braddy, 2006). Following the recommendation by Keith (2015), two perspectives were jointly evaluated: (a) the traditional perspective based on $\Delta\chi^2$, and (b) the practical perspective based on differences in the comparative fix index CFI (ΔCFI). Comparatively, the $\Delta\chi^2$ test is known to be oversensitive to the sample size and discrepancies from normality (Kline, 2010; West et al., 1995). Cheung and Rensvold (2002) recommended ΔCFI as superior to $\Delta\chi^2$ for its independence in model complexity, sample size, and overall fit measures. "A value of ΔCFI smaller than or equal to −.01 indicates that the null hypothesis of invariance should not be rejected" (Cheung & Rensvold, 2002, p. 251). An absolute ΔCFI value higher than .01 (i.e., $|\Delta CFI| > .01$) was proposed as an indicator of a meaningful fall in fit. Given the large sample sizes, large modeled variables, and the number of comparisons being made in this study, we decided to evaluate the invariance by $\Delta\chi^2$ and ΔCFI jointly to secure meaningfulness and prevent any unnecessary oversensitivity. The criterion for rejecting the null hypothesis of invariance was set as a *p* value of less than .001 for the $\Delta\chi^2$ test and an absolute ΔCFI value higher than .01.

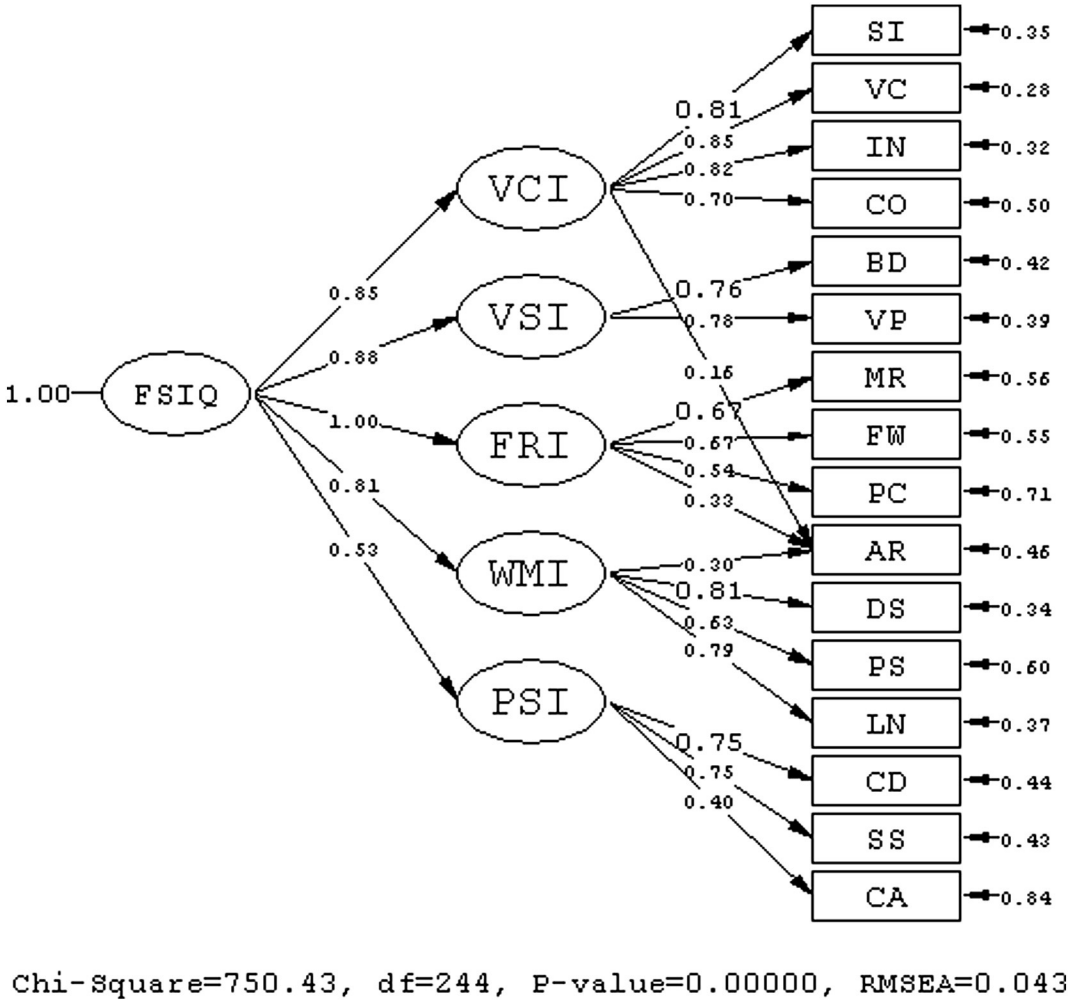

Fig. 1. Final standardized estimations of both genders on the 16 subtests (Model 7 in Table 1).

3. Results

Table 1 lists all steps in the invariance analyses. The baseline model fit was first checked for each sample. The model fits each datum well, suggesting that the following invariance verification was meaningful. Variance–covariance matrices were constrained to be equal across groups (Model 1). This constrained model fits the data well (CFI = 1.00; RMSEA = .011), suggesting fairly invariant WISC-V subtest covariance patterns in children. Because equality of variance–covariance matrices between genders is supported, the WISC-V generally measures the same constructs between genders. Because any factor structure is derived from these variance–covariance matrices, this result revealed that the WISC-V factor structure between males and females should be similar.

Table 1
Multi-sample goodness-of-fit indices for the WISC—V 16 subtests.

Model	χ^2	*df*	CFI	RMSEA	RMSEA 90%CI	SRMR	AIC	aBIC	Model comparison	ΔCFI	$\Delta\chi^2$	Δdf	*p*
Phase I: Baseline model fit for each group													
Male (n = 1099)	223.30	97	.99	.034	.029–.040	.027	301.30						
Female (n = 1101)	186.89	97	1.00	.029	.023–.035	.025	264.89						
Phase II: Measurement Invariance across groups													
Model 1 Equality of variance–covariance matrices	154.64	136	1.00	.011	.000–.019	.038	426.64	769.23					
Model 2 configural	410.19	194	.99	.032	.028–.036	.025	630.19	907.29	–	–	–	–	–
Model 3 first-order loadings	428.14	207	.99	.031	.027–.035	.029	622.14	866.49	3 vs. 2	0	17.95	13	.159
Model 4 first-order loadings and subtest intercepts	702.87	218	.99	.045	.041–.049	.032	874.87	1091.51	4 vs. 3	0	274.73	11	.000
Model 5 first-order loadings, subtest intercepts, and subtest residual variances	738.36	234	.99	.044	.041–.048	.034	878.36	1054.69	5 vs. 4	0	35.49	16	.003
Model 6 first-order loadings, subtest intercepts, residual variances, and second-order loadings	742.98	239	.99	.044	.040–.047	.043	872.98	1036.72	6 vs. 5	0	4.62	5	.463
Model 7 first-order loadings, subtest intercepts, residual variances, second-order loadings, and disturbances of first-order factors	750.43	244	.99	.043	.040–.047	.045	870.43	1021.57	7 vs. 6	0	7.45	5	.189

When testing nested models, first, the configural model (Model 2) provided an acceptable fit to the data. Males and females shared the same WISC-V first- and second-order five-factor patterns and the corresponding subtests loaded on the same factors. With the factor pattern established, we imposed cross-group constraints on the first-order factor loadings (Model 3). There was no deterioration of fit with these constraints by both $\Delta\chi^2$ and ΔCFI, implying that the subtests measure the same latent factors in both groups. Next, we constrained the subtest intercepts to be equal (Model 4). To identify this model properly, we fixed the means of the first-order factors in the male group to zero, but freed those in the female group. Thus, the factor means for the female group represent the mean differences. The addition of subtest intercepts constraints reduced the fit according to $\Delta\chi^2$, but not according to ΔCFI. The ΔCFI value was 0, implying that the subtest intercepts are the same in both groups. Next, when the subtest residuals were constrained to be equal across groups (Model 5), there was no deterioration of fit with these constraints. When structural parameters (second-order loadings and first-order unique variances) were constrained as to be equal between groups in steps (Models 6 and 7), again, there was no result in practical deterioration of fit by both $\Delta\chi^2$ and ΔCFI.

Because of the complexity of the model and the strictness of the test, we concluded that the WISC-V exhibits acceptable levels of invariance among five factors between the male and female groups. Differences in subtest scores on the WISC-V are generally caused by latent constructs, and the test is not biased based on the gender status. We fixed the means of the five latent factors in the male group to zero, and the non-standardized latent means for the Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed factors in the female group were estimated freely as −0.18, −0.20, −0.06, 0.14, and 1.04, respectively. This finding suggests that male and female children scored closely on most of the WISC-V underlying first-order factors. The largest discrepancy emerged for the Processing Speed factor, with a mean difference over one-third standard deviation.

Standardized estimates based on Model 7 for both groups are shown in Fig. 1. All 16 subtests loaded strongly on the corresponding factors. Consistent with the literature, Arithmetic was confirmed as a mixed measure of the Fluid Reasoning, Working Memory, and Verbal Comprehension factors (factor loadings were .33, .30, and .16, respectively). Across all five first-order factors, fluid reasoning had the highest *g* loading (1.00). All parameter estimates were theoretically reasonable. Most importantly, these estimates were found invariant between genders.

Table 2 lists *g* loadings for each of the WISC-V subtests. For children from both genders, the subtests with top *g* loadings are: Vocabulary, Arithmetic, Information, Similarities, and Visual Puzzles.

Table 2
Loadings of WISC—V subtests on the second-order *g* factor.

Subtest	*g*-Loading
Vocabulary (VC)	.72
Arithmetic (AR)	.71
Information (IN)	.70
Similarities (SI)	.69
Visual Puzzles (VP)	.69
Matrix Reasoning (MR)	.67
Block Design (BD)	.67
Figure Weights (FW)	.67
Digit Span (DS)	.66
Letter–Number Sequencing (LN)	.64
Comprehension (CO)	.60
Picture Concepts (PC)	.54
Picture Span (PS)	.51
Coding (CD)	.40
Symbol Search (SS)	.40
Cancellation (CA)	.21

4. Discussion

We conducted this study to determine the invariance of WISC-V constructs across large samples of male and female children. This study is valuable, as it is the first to evaluate the gender invariance of the newly published WISC-V five-factor interpretive approach.

The first and most critical set of findings is that the five-factor model fits the data from both genders well. This model demonstrated full factorial invariance between genders. Across genders the WISC-V subtests generally demonstrate the same underlying theoretical latent constructs, the same strength of relationships among factors and subtests, the same validity of each first-order factor, and the same communalities. Invariant results provide evidence that WISC-V index scores and subtests have the same meaning for both genders, WISC-V results for males and females can be interpreted in the same way, and that meaningful comparisons between genders can be made.

The second set of findings concerns the verification of multiple abilities, as required by some subtests, as invariant between genders. Many previous studies reported the mixed loadings of the Arithmetic subtest (Chen, Keith, Chen, & Chang, 2009; Weiss et al., 2013). Current findings further demonstrated that these previously identified cross-loadings exist for both male and female children. Regardless of gender then, when interpreting WISC-V results, performance on the Arithmetic subtest should be considered to be influenced mainly by fluid reasoning (.33) and working memory (.30) abilities, also to some extent by verbal comprehension (.16) ability.

A third set of major findings is that the Fluid Reasoning factor had a standardized loading of 1.00 on the second-order *g* factor. In the literature, there are considerable reports suggesting that fluid reasoning factors often show *g* loadings approaching or even reaching unity (Bickley, Keith, & Wolfle, 1995; Gutafsson, 1984; Keith et al., 2006). Once again, fluid reasoning is demonstrated to be the cornerstone of human cognition. Among all subtests, Vocabulary had the highest *g*-loading (0.72), followed by Arithmetic (0.71), Information (0.70), and Similarities (.69). The Cancellation subtest had the lowest *g*-loading (0.21). These findings are similar to those reported for the WISC-IV (Keith et al., 2006). The new WISC-V subtests, Visual Puzzles and Figure Weights, were found to show high *g*-loadings for both genders. Results indicate that these new subtests make profound contributions to the WISC-V latent construct.

Finally, our results reveal that male and female children scored closely on the WISC-V latent factors. Comparatively, male children seemed to perform slightly better on visual spatial (with a small effect of latent mean difference of .07 of a standard deviation), and females seemed to perform better on processing speed (with a large effect of latent mean difference of .35 of a standard deviation). This finding is consistent with many reports in the literature (Chen, Chen, Chang, Lee, & Chen, 2010; Chen, Keith, et al., 2010; Gallagher & Kaufman, 2005).

In conclusion, our findings yielded strong support for the invariant WISC-V factor structure between genders. The meaning of each WISC-V subtest and factor-based composite is generally identical for each gender. WISC-V scores for males and females can be interpreted in the same way.

We recommend that future validity evidence be accumulated continuously. Invariant meaning for children in other subpopulations (e.g., ages, clinical groups, or cultures) be explored, and studies based on clinical performance or diagnostic differentiation be conducted to provide more evidence of validity and increase our understanding of how the WISC-V functions in various relevant groups of children.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Archer, R. P., Buffington-Vollum, J. K., Stredny, & Handel, R. W. (2006). A survey of test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*, 64–94. http://dx.doi.org/10.1207/s15327752jpa8701_07.

Atienza, F. L., Balaguer, I., & Garcia-Merita, M. L. (2003). Satisfaction with life scale: Analysis of factorial invariance across sexes. *Personality and Individual Differences*, *35*, 1255–1260.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, *20*, 309–328.

Bowden, S. C. (2013). Theoretical convergence in assessment of cognition. *Journal of Psychoeducational Assessment*, *31*(2), 148–156.

Byrne, B. M., Baron, P., & Campbell, T. L. (1993). Measuring adolescent depression: Factorial validity and invariance of the Beck Depression Inventory across gender. *Journal of Research on Adolescence*, *3*, 127–143.

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*(2), 287–321.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, *31*, 141–154.

Canivez, G. L., & Watkins, M. W. (2015). Review of the Wechsler Intelligence Scale for Children-Fifth Edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Raiford, & D. L. Coalson (Eds.), *Intelligence testing with the WISC-V*. Hoboken, NJ: Wiley (in press).

Chen, H., Chen, M., Chang, T., Lee, Y., & Chen, H. P. (2010a). Gender reality on multidomains of school-age children in Taiwan: A developmental approach. *Personality and Individual Differences*, *48*(4), 475–480.

Chen, H., Hung, L. Y., Chen, Y., Zhu, J., & Keith, T. (2015). Validation of Taiwan WISC-IV by four- and five-factor interpretative approaches. *Bulletin of Special Education* (in press).

Chen, H., Keith, T., Chen, Y., & Chang, B. (2009). What does the WISC-IV measure?: Validation of the scoring and CHC-based interpretative approaches. *Journal of Research in Education Sciences*, *54*(3), 85–108.

Chen, H., Keith, T., Weiss, L., Zhu, J., & Li, Y. (2010b). Testing for multigroup invariance of second-order WISC-IV structure across China, Hong Kong, Macau, and Taiwan. *Personality and Individual Differences*, *49*(7), 677–682. http://dx.doi.org/10.1016/j.paid.2010.06.004.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*(3), 471–492.

Chen, H., & Zhu, J. (2008). Factor invariance between genders of the Wechsler Intelligence Scale for Children-Fourth Edition. *Personality and Individual Differences*, *45*(3), 260–266. http://dx.doi.org/10.1016/j.paid.2008.04.008.

Chen, H., & Zhu, J. (2012). Measurement invariance of WISC-IV across normative and clinical samples. *Personality and Individual Differences*, *52*(2), 161–166. http://dx.doi.org/10.1016/j.paid.2011.10.006.

Cheng, C. H. K., & Watkins, D. (2000). Age and gender invariance of self-concept factor structure: An investigation of a newly developed Chinese self-concept instrument. *International Journal of Psychology*, *35*(5), 186–193.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin*, *95*, 134–135.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19–29.

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC-IV assessment*. New York, NY: John Wiley & Sons, Inc.

Gallagher, A. M., & Kaufman, J. C. (Eds.). (2005). *Gender differences in mathematics: An integrative psychological approach*. Cambridge, UK: Cambridge University Press.

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., & Saklofske, D. H. (Eds.). (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC-III*. San Diego, CA: Academic Press.

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Gutafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179–203.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.

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

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8 statistical program*. Lincolnwood, IL: Scientific Software.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structure equation modeling* (2nd ed.). NY: Routledge.

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

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th rev. ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391–410.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2006). The utility of alternative fit indices in tests of measurement invariance. *Academy of Management Annual Meeting Proceedings, pB1*.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*(1), 93–115.

Prifitera, A., Saklofske, D. H., & Weiss, L. G. (Eds.). (2005). *WISC-IV clinical use and interpretation: Scientist–Practitioner perspectives*. San Diego, CA: Academic Press.

Prifitera, A., Saklofske, D. H., & Weiss, L. G. (Eds.). (2008). *WISC-IV clinical assessment and intervention 2e*. San Diego, CA: Academic Press.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: a survey of INS, NAN, and APA division 40 members. *Archives of Clinical Neuropsychology*, *20*, 33–65. http://dx.doi.org/10.1016/j.acn.2004.02.005.

Richardson, C. E., Huan, J., Ege, E., Suh, H., & Rice, K. G. (2014). Refining the measurement of maximization: Gender invariance and relation to psychological well-being. *Personality & Individual Differences*, *70*, 229–234.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*, 403–418.

Rusticus, S. A., & Hubley, A. M. (2006). Measurement invariance of the Multidimensional Body-Self Relations Questionnaire: Can we compare across age and gender? *Sex Roles*, *55*, 827–842.

Sattler, J. M., & Dumont, R. (2004). *Assessment of children: WISC-IV and WPPSI-III supplement*. Le Mesa, CA: Author.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158.

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*(1), 4–69.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children-Fourth Edition. Administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children-Fifth Edition. Administration and scoring manual*. Bloomington, MN: Pearson.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children-Fifth Edition. Technical and interpretive manual*. Bloomington, MN: Pearson.

Weiss, L. G., Keith, T., Zhu, J., & Chen, H. (2013). WISC-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*(2), 114–131.

Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (Eds.). (2008). *WISC-IV advanced clinical interpretation*. San Diego, CA: Academic Press.

West, S. G., Finch, J. P., & Curran, P. J. (1995). Structural equation models with nonnormal variables. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39–47. http://dx.doi.org/10.1111/j.1745-3992. 2010. 00182.x.

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* (pp. 281–324). Washington: American Psychological Association.