Personality and Individual Differences 45 (2008) 260–266

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Factor invariance between genders of the Wechsler Intelligence Scale for Children – Fourth edition

Hsinyi Chen [a,*], Jianjun Zhu [b]

[a] Department of Special Education, National Taiwan Normal University, Taipei, Taiwan, ROC
[b] Clinical Products, The Psychological Corporation, San Antonio, TX, USA

ARTICLE INFO

Article history:
Received 22 January 2008
Received in revised form 29 March 2008
Accepted 8 April 2008
Available online 19 May 2008

Keywords:
WISC-IV
Multi-group SEM
Gender invariance

ABSTRACT

The purpose of this study was to test measurement invariance of the WISC-IV factorial structure between genders. Both 10 and 15 subtest sets were tested on a nationally-representative sample of 2200 children (males $N = 1100$; females $N = 1100$). Results from multisample analyses supported partial measurement invariance. Except for a few subtle discrepancies involving correlated error terms, the hypothesized model described the data for both sexes. Overall factor patterns, loadings, unique variances, and factor covariances of the WISC-IV generally did not vary with gender.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Wechsler tests are among the most widely used intelligence tests in the world. Roughly, 20 countries have standardized these tests thus far (Camara, Nathan, & Puente, 2000; Georgas, Weiss, van de Vijver, & Saklofske, 2003). In 1991, the Wechsler Intelligence Scale for Children – Third Edition (WISC-III; Wechsler, 1991) introduced a four-factor solution as an alternative to the traditional Verbal and Performance IQ scores. This four-factor model is recognized as more in line with contemporary research on intellectual constructs and is extensively cross-validated in a variety of samples (Donders & Warschausky, 1996; Keith & Witta, 1997; Konold, Kush, & Canivez, 1997; Roid & Worral, 1997; Tupa, Wright, & Fristad, 1997). The recently published fourth edition of the Wechsler Intelligence Scale for Children (WISC-IV; Wechsler, 2003) refined these four factors into Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. Recent studies examined the construct validity for normative samples (Keith, Fine, Taub, Reynolds, & Kranzler, 2006; Watkins, 2006) and for referred students (Watkins, Wilson, Kotz, Carbone, & Babula, 2006). The fundamental issue of gender invariance, however, remains to be answered.

Invariance is a critical property for any measure (Drasgow, 1984, 1987; Rock, Werts, & Flaugher, 1978). It assumes that the test measures the same constructs in different subgroups or assessment occasions. A lack of evidence for measurement invariance for a particular construct obviates the ability of the measure to be used in comparisons among groups on that construct (Horn & McArdle, 1992; Vandenberg & Lance, 2000).

Among all possible subgroup classifications, gender invariance is recognized as the one fundamental issue for measurements in domains such as depression (Byrne, Baron, & Balev, 1996; Byrne, Baron, & Campbell, 1993, 1994), satisfaction with life (Atienza, Balaguer, & Garcia-Merita, 2003; Shevlin, Brunsden, & Miles, 1998; Wu & Yao, 2006), meaning in life (Reker, 2005), self-concept (Cheng & Watkins, 2000; Marsh, 1993; Yin & Fan, 2003), body-self relationship (Rusticus & Hubley, 2006), positive and negative affect (Mackinnon et al., 1999), anxiety sensitivity (Dehon, Weems, Stickle, Costa, & Berman, 2005), and decisional balance (Ward, Velicer, Rossi, Fava, & Prochaska, 2004). Gender invariance certainly is an essential issue pertaining to WISC-IV. Data from males and females are usually combined when substantive applied studies are conducted empirically. There is considerable interest in gender differences in cognitive abilities (Hyde, Fennema, & Lamon, 1990; Linn & Petersen, 1985; Rosenthal & Rubin, 1982; Wang & Maxey, 1995). Implicit in this common practice of lumping gender data is the assumption that WISC-IV subtests and index scores have the same meaning for both sexes. According to standard 7.8 of the "Standards for Educational and Psychological Testing" (American Educational Research Association & National Council on Measurement in Education, 1999), "Comparisons across groups are only meaningful if scores have comparable meaning across groups. The standard is intended applicable to settings where scores are implicitly or explicitly presented as comparable in meaning across groups (p. 83)". Furthermore, Millsap and Kwok (2004) suggested that lack of invariance may affect accuracy of selection. Since considerable attention is given to fairness in

* Corresponding author. Tel.: +886 02 23568901x211; fax: +886 02 23413061.
E-mail address: hsinyi@ntnu.edu.tw (H. Chen).

0191-8869/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2008.04.008

selection between genders (Willingham & Cole, 1997) and since Wechsler composite scores are frequently used in formal identification scenarios, it is important to test the assumptions of gender invariance in WISC-IV.

2. Method

2.1. Participants

We analyzed the WISC-IV standardization responses from 2200 children (males $N = 1100$; females $N = 1100$). 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 2000 United States Census on region, gender, parent educational level, and ethnicity. A detailed description of this sample is provided in the WISC-IV manual (Wechsler, 2003).

2.2. Instrumentation

The WISC-IV has 10 core subtests and five supplemental subtests. The 10 core subtests are Similarities (SI), Vocabulary (VC), Comprehension (CO), Block Design (BD), Picture Concepts (PS), Matrix Reasoning (MR), Digit Span (DS), Letter–Number Sequence (LNS), Coding (CD), and Symbol Search (SS). The five supplemental subtests are Information (IN), Word Reasoning (WC), Picture Completion (PC), Arithmetic (AR), and Cancellation (CA).

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). Because the WISC-IV includes a core version with 10 subtests and a core plus supplemental version, involving all 15 subtests, we used the subtest scaled scores to study the invariance issue of both versions.

Prior to invariance analysis, we separately tested the corresponding four-factor baseline model for males and females. The structure reported in the WISC-IV manual (Wechsler, 2003) was used as the hypothesized baseline model. For the 10-subtest version, the baseline model specified 3 Verbal Comprehension subtests (SI, VC, CO) on the first factor, 3 Perceptual Reasoning subtests (BD, PS, MR) on the second factor, 2 Working Memory subtests (DS, LNS) on the third factor, and 2 Processing Speed subtests (CD, SS) on the fourth factor. For the 15-subtest version, there were 2 extra Verbal Comprehension subtests (IN, WC), 1 extra Perceptual Reasoning subtest (PC), 1 extra Working Memory subtest (AR), and 1 extra Processing Speed subtest (CA). No correlated residuals were specified.

For each baseline model, we tested for invariance on four levels of nested models. Each level had more constraints than the previous one (Meredith, 1993). The first and weakest level was configural invariance, it assumed the overall factor pattern was the same between genders. The second level is testing for weak factorial invariance, also called metric invariance. This model requires the magnitude of the factor loadings be the same between genders ($\Lambda_m = \Lambda_f$). The third stage tests unique variance invariance ($\Theta_m = \Theta_f$), which examines whether the test measures the same construct with similar accuracy. Finally, in the most restricted model, the factor covariances were all constrained to be equal between genders ($\Phi_m = \Phi_f$).

All models were tested using covariance matrices. The scale of latent factors was defined by fixing factor variance to one. Criteria were evaluated jointly to assess overall model fit (Bentler & Bonett, 1980; Marsh, Balla, & McDonald, 1988). These included weighted least squares χ^2, χ^2 to df ratio, goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), comparative fit index (CFI), root-mean-square-error of approximation (RMSEA), and standardized root mean square residual (SRMR). A value of .90 served as the rule-of-thumb indication of good fit for all fit indices ranging from zero to 1, with 1 indicating a perfect fit (Hoyle & Panter, 1995; Kline, 2005). Values close to 2.0 or 3.0 were considered fit for the χ^2 to df ratio (Bollen, 1989). A RMSEA less than .05 corresponded to a "good" fit and with .08 considered an "acceptable" fit (McDonald & Ho, 2002). Finally, SRMR values less than .08 are considered acceptable (Hu & Bentler, 1999).

During each step of the analyses, the chi square difference ($\Delta\chi^2$) was tested between nested models, and suggestions regarding to partial measurement invariance (Byrne, Shavelson, & Muthen, 1989; Byrne & Watkins, 2003) was carefully considered and followed. If inadequate fit was detected, fit in the model was improved by including additional parameters identified by the modification index (MI) provided by LISREL. Re-parameterization was examined carefully for meaningfulness.

3. Results

3.1. Normality checking

Skewness and kurtosis for each subtest by gender are presented in Table 1. Also presented is the D'agostino-Pearson omnibus K^2 test (D'Agostino, Belanger, & D'Agostino, 1990; D'Agostino and Pearson, 1973) for normality testing.

Skewness ranged from −.55 to .16 for males and from −.86 to .10 for females; kurtosis ranged from −.26 to .78 for males and from −.35 to 1.35 for females. Roughly, half of the K^2 statistics were statistically significant. To deal with non-normal data with such a large sample size, the generally weighted least-squares estimation (WLS) (Browne, 1984; Jöreskog & Sörbom, 1996) was selected as the procedures for model estimation.

3.2. The 10 core subtests

3.2.1. Baseline model Identification

As indicated by all goodness-of-fit indexes reported in Table 2, the initially hypothesized four-factor model fit equally well for both male and female groups. Since goodness-of-fit values were

Table 1
Normality of subtests in the Wechsler Intelligence Scale for Children – Fourth Edition

	Males ($N = 1100$)			Females ($N = 1100$)		
Subtest	Sk	Ku	K^2	Sk	Ku	K^2
SI	−.23	−.19	11.23**	−.05	−.35	8.64*
VC	−.36	.21	25.03**	−.35	.16	23.22**
CO	−.23	.40	15.32**	−.21	.45	14.66**
IN	−.02	−.01	.07	.06	.22	2.83
WC	−.15	−.14	5.31	−.11	.09	2.77
BD	−.07	−.04	.88	−.09	−.06	1.61
PS	−.32	.35	22.52**	−.37	.18	25.61**
MR	.16	.11	5.10	.10	.10	2.27
PC	−.26	.19	13.75**	−.16	.15	5.63
DS	.09	.48	9.17*	−.14	.26	6.48*
LNS	−.55	.78	64.89**	−.86	1.35	137.76**
AR	−.02	−.15	.77	−.00	.28	2.23
CD	.11	.17	3.58	−.16	.33	8.56*
SS	−.29	.48	22.89**	−.45	.84	52.56**
CA	−.08	−.26	4.97	−.05	−.02	.49

Note: Sk, Skewness; Ku, Kurtosis; K^2, D'agostino-Pearson omnibus test for normality; SI, Similarities; VC, Vocabulary; CO, Comprehension; IN, Information; WC, Word Reasoning; BD, Block Design; PS, Picture Concepts; MR, Matrix Reasoning; PC, Picture Completion; DS, Digit Span; LNS, Letter–Number Sequence; AR, Arithmetic; CD, Coding; SS, Symbol Search; CA, Cancellation; * $p < .05$. ** $p < .01$.

متن کامل مقاله

دریافت فوری

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات