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 1 June 2015

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.