Personality and Individual Differences 49 (2010) 677–682

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Testing for multigroup invariance of second-order WISC-IV structure across China, Hong Kong, Macau, and Taiwan

Hsinyi Chen [a,*], Timothy Z. Keith [b], Larry Weiss [c], Jianjun Zhu [c], YuQiu Li [d]

[a] Department of Special Education, National Taiwan Normal University, Taipei, Taiwan, ROC
[b] Department of Educational Psychology, The University of Texas at Austin, Austin, TX, USA
[c] Clinical Assessment, Pearson, San Antonio, TX, USA
[d] College of Education, Beijing Normal University at Zhuhai, China

ARTICLE INFO

Article history:
Received 26 March 2010
Received in revised form 27 May 2010
Accepted 6 June 2010

Keywords:
WISC-IV
Cultural invariance
China
Hong Kong
Macau
Taiwan

ABSTRACT

We tested measurement invariance of the WISC-IV second-order factorial structure across China, Hong Kong, Macau, and Taiwan. Both sets of 10 and 14 subtests were tested on standardization samples of children aged 6–16. Results from multi-sample analyses supported measurement invariance, the hypothesized second-order factor model well described data from all four cultures. Overall factor patterns, first- and second-order factor loadings, *g* variance, residual variances of measured variables, and disturbances of first-order factors of the WISC-IV were invariant among these four cultures.

1. Introduction

Wechsler tests are among the most widely used intelligence tests in the world. Roughly 20 countries have standardized these tests thus far (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 traditional Verbal and Performance IQ scores. This four-factor model is recognized as more in line with present-age research on intellectual constructs and is extensively cross-validated in a variety of samples (Donders & Warschausky, 1996; Konold, Kush, & Canivez, 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 have examined its validity for normative samples (Chen, Keith, Chen, & Chang, 2009; Keith, Fine, Taub, Reynolds, & Kranzler, 2006) and for referred students (Watkins, Wilson, Kotz, Carbone, & Babula, 2006). Although the four-factor solution was found to fit well in different nations, it is unknown whether it shows measurement and structural invariance across cultures.

Invariance is a critical property for any measure (Drasgow, 1984, 1987). It assumes that the test measures the same constructs in different groups. Meaningful comparisons of statistics such as regression coefficients and means can only be made if the measures are comparable across different groups (Chen, Sousa, & West, 2005). 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 (Vandenberg & Lance, 2000). According to standard 7.8 of the "Standards for Educational and Psychological Testing" (American Educational Research Association, American Psychological 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)".

Cultural invariance is an essential issue pertaining to WISC-IV. There is considerable interest in cross-cultural studies of cognitive processes in different cultures. Without evidence of the WISC-IV measurement invariance, interpretation of WISC-IV-based universals and variations in cognitive process is impossible. Implicit in this common practice of cross-cultural studies of intelligence is the assumption that subtests and index scores have the same meaning for different cultures. Thus, the measurement invariance must be examined prior to any interpretations of cultural differences.

* Corresponding author. Tel.: +886 (02)7734 5011; fax: +886 (02)23413061.
E-mail address: hsinyi@ntnu.edu.tw (H. Chen).

0191-8869/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2010.06.004

Our purpose was to test measurement invariance of the newly adapted WISC-IV across China, Hong Kong, Macau, and Taiwan. This was the first time in decades that the same Wechsler version was compared across these four cultures. Our evaluation of the WISC-IV invariance addresses an important issue.

2. Method

2.1. Participants

The data sets we analyzed were from the standardization studies of the WISC-IV in China (n = 1100), Hong Kong (n = 550), Macau (n = 298), and Taiwan (n = 968). All samples were selected to match recent censuses for major demographics such as region, gender, parent educational level, and ethnicity. Each representative sample was divided into 11 age groups from ages 6 to 16, with a balanced number of children in each age group. Detailed descriptions of each of these norming samples are provided in the corresponding WISC-IV manuals (Wechsler, 2007, 2008, 2009, 2010).

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 (LN), Coding (CD), and Symbol Search (SS). The five supplemental subtests are Information (IN), Word Reasoning (WC), Picture Completion (PC), Arithmetic (AR), and Cancellation (CA). The Word Reasoning subtest was not adapted in any of these four cultures. Therefore, a total of 14 subtests were analyzed in this present work.

2.3. Analysis of the data

Contemporary studies of intelligence generally agree upon a hierarchical model of cognitive abilities. General intelligence (g) tends to emerge whenever a sufficient number of cognitively complex variables are analyzed (Carroll, 1993). Prior to invariance analysis, we separately tested the second-order baseline model for each of these four cultures. The scoring structure reported in the WISC-IV manual (Wechsler, 2003) was used as the hypothesized baseline model. With the 10 core subtests and four supplemental subtests divided into four factors: Verbal Comprehension (includes subtests SI, VC, CO, and IN), Perceptual Reasoning (includes subtests BD, PS, MR, and PC), Working Memory (includes subtests DS, LN, and AR), and Processing Speed (includes subtests CD, SS, and CA). Superior to these four first-order factors is the higher order factor g, no correlated residuals were specified. Compared to a correlated four-factor model, the second-order factor model is known to be more parsimonious, more consistent with contemporary theory, and more consistent with the structure of the WISC-IV. Tests of invariance in both the 10- and 14-subtest sets were based on the analysis of covariance structure models using LISREL 8.8 (Jöreskog & Sörbom, 2006).

For each baseline model, we first test equality of variance and covariance matrices, to investigate whether the WISC-IV measures the same construct across cultures. If so, we further tested six levels of nested models to identify the nature of this construct. Each level had more constraints than the previous one (Byrne & Stewart, 2006; Chen et al., 2005). The initial and weakest level was configural invariance, it assumed that the same number of factors and overall factor pattern was the same across cultures. The second and third levels were factor loading invariance, also called metric invariance. These models required the magnitude of first-order and second-order factor loadings be the same across groups. When the factor loadings are equal across groups, this means that the unit of measurement is identical across groups. That means, for example that, a 10-point increase in latent Verbal Comprehension ability would result in the same increase in Vocabulary subtest performance for one culture as for another. Likewise, with invariance of second-order loadings, a 10-point increase in latent g would result in the same increase in latent Verbal Comprehension ability across all cultures. In other words, with factor loading invariance, the same measurement scale is used across cultures. 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), we tested invariance of residuals (a combination of random measurement error and subtest-specific unique variance). We also tested for the invariance of the disturbances (factor unique variance) of the first-order factors. Finally, we tested the most restricted model, where the highest order g factor variances were constrained as equal across cultures.

All models were tested using covariance matrices. Scale of latent factors was identified by fixing a factor loading for each factor to one. Criteria were evaluated jointly to assess overall model fit (Hu & Bentler, 1998; Kline,2005), including weighted least squares χ^2, χ^2 to df ratio, comparative fit index (CFI), root-mean-square-error of approximation (RMSEA), standardized root mean square residual (SRMR), and Akaike information criterion (AIC). A value of .95 served as the rule-of-thumb cut point of acceptable fit for CFI (Hu & Bentler, 1999). Values close to 2.0 or 3.0 were considered good fits for the χ^2 to df ratio (Bollen, 1989). An RMSEA less than .05 corresponded to a good fit and with .08 considered an acceptable fit (McDonald & Ho, 2002). For completeness, we included the 90% confidence interval for the RMSEA. Finally, SRMR values less than .08 were considered acceptable (Hu & Bentler, 1999). The AIC was reported for testing non-nested rival models (Kaplan, 2000), with smaller AIC values indicating a better fit.

To evaluate invariance of competing models, traditionally, the Likelihood-ratio test, also known as the chi-square difference test ($\Delta\chi^2$) is used to test nested models (Loehlin, 2004). Since this test is sensitive to sample size, size of the correlations, and moderate discrepancies from normality (Kline, 2005; West, Finch, & Curran, 1995), we followed the recommendation of Cheung and Rensvold (2002) and added ΔCFI. This test is superior to $\Delta\chi^2$ as a test of invariance because it is independent of both model complexity and sample size, and is not correlated with the overall fit measures. "A value of ΔCFI smaller than or equal to −.01 indicates that the null hypothesis of invariance should not be rejected (p. 251)". Both $\Delta\chi^2$ and ΔCFI were considered in the invariance evaluation process.

3. Results

3.1. Normality checking

Skewness and kurtosis for each subtest by culture were presented in Table 1. We tested for normality of each subtest with the D'Agostino–Pearson omnibus K^2 test (D'Agostino, Belanger, & D'Agostino, 1990; D'Agostino and Pearson, 1973).

Across all four cultures, Skewness ranged from −.65 to .39, and kurtosis ranged from −.39 to 1.82. Although many K^2 statistics were statistically significant, the effects were not large. According to a rule-of-thumb by West et al. (1995), maximum Likelihood estimation seems to work well with skewness less than 2 and kurtosis less than 7. This method is also known for its robustness and sensitivity to incorrectly specified models (Hu & Bentler, 1998).