Learning and Individual Differences 45 (2016) 307–312

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Are there sex differences in the Wechsler Intelligence Scale for Children — Forth Edition?

Lina Pezzuti *, Arturo Orsini

Department of Clinical Psychology, Sapienza University of Rome, Via dei Marsi 78, 00185 Rome, Italy

ARTICLE INFO

Article history:
Received 11 September 2014
Received in revised form 30 November 2015
Accepted 27 December 2015

Keywords:
Sex differences
Intelligence
Children
Adolescent
WISC-IV

ABSTRACT

In the Wechsler Intelligence Scale for Children — Forth Edition (WISC-IV), the manual reports several confirmatory factor analyses in support of the instrument's latent factor structure, but no information about eventually sex differences. The present study aims to investigate the factorial invariance and factor mean differences of the hierarchical model of WISC-IV, between Italian males and females. The overall results from this study generally support both configural and factorial invariance of the WISC-IV when the 10 core subtest are administered, so the second-order WISC-IV structure is equivalent across the females and males. In estimating latent mean differences on the Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed and General Intelligence, the results was that the males were higher in the Verbal Comprehension and the females were higher in the Processing Speed. These results were partially confirmed by results on observed means differences. No significant differences in the latent factor means and observed means of Perceptual Reasoning, Working Memory and General Intelligence were found.

1. Introduction

The Wechsler Intelligence Scale for Children — Forth Edition (WISC-IV, Wechsler, 2003) represents the latest edition of this intelligence battery. It is the most widely used test of children and adolescent intelligence. In this edition the structure of the instrument has changed, and some subtests have been added and others deleted. This edition refined the four-factor solution (four Indexes), almost introduced with WISC-III (Wechsler, 1991), into Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed. Thus, the WISC-IV allows for better discrimination between abilities on the aggregate level than its predecessors. All standardizations of the Wechsler scales do not provide sex-differentiated norms: this means that it is assumed that there are no sex differences in the level of intelligence. In particular, implicit is the assumption that WISC-IV subtests and Index scores have the same meaning for both sexes.

The findings on cognitive sex differences are still controversial. There are authors argue that are no or only minimal sex differences in general cognitive performance (Colom, Juan-Espinosa, Abad, & Garcia, 2000; Halper & LaMay, 2000; Hines, 2007; Jensen & Johnson, 1994). On the contrary authors describe a superiority of males in Full Scale IQ (FSIQ) on WISC-R (Born & Lynn, 1994; Lynn & Mulhern, 1991; Lynn, Raine, Venables, Mednick, & Irwing, 2005).

Historically, sex differences are reported in task-specific test performance: for example, a superiority of females in verbal abilities and verbal memory, and a superiority of males in spatial tasks and mathematical reasoning (Keith, Fine, Taub, Reynolds, & Kranzler, 2006; Keith, Reynolds, Patel, & Ridley, 2008; Nyborg, 2005; Reynolds, Keith, Ridley, & Patel, 2008).

According the studies on WISC-IV, the results of Chen and Zhu (2008) from multisample analyses, the hypothesised WISC-IV four-factor 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 sex.

Goldbeck, Daseking, Hellwig-Brida, Waldmann, and Petermann (2010) reported for all age groups there were no sex effects in the Full-Scale IQ, but sex effects favouring males in the Verbal Comprehension Index and Perceptual Reasoning Index. On the contrary, the girls scored higher than boys in the Processing Speed Index.

The aim of the present study is twofold: (1) to test for measurement equivalence of the WISC-IV across males and females groups based on one hierarchical factorial structure that include one higher-order factor of General Intelligence and four lower-order factors of Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed: this hypothesised structure of WISC-IV derive from the structure

* Corresponding author.
E-mail addresses: lina.pezzuti@uniroma1.it (L. Pezzuti), arturo.orsini@uniroma1.it (A. Orsini).

http://dx.doi.org/10.1016/j.lindif.2015.12.024
1041-6080/

reported in the WISC-IV manual (Wechsler, 2003); (2) to test for differences in the means (or levels) of these higher- and lower-order factors across these samples.

2. Method

2.1. Participants

The data used in the present paper involving the Italian sample of WISC-IV standardization (Orsini, Pezzuti, & Picone, 2012; Wechsler, 2012). The Italian standardization sample comprises 2200 participants (1100 females and 1100 males) divided into 11 age groups from 6 years old to 16 years old. Each age group therefore comprises 200 participants (100 females and 100 males). The sample is representative of the Italian population according to level of parental education. Participants were tested in their school in individual sessions lasting between 45 and 70 min in a quiet room away from the classroom. The informed consent of the parents were asked.

2.2. Instrumentation

The WISC-IV has 10 core subtests (Similarity, Vocabulary, Comprehension, Block Design, Picture Concepts, Matrix Reasoning, Digit Span, Letter-Number Sequence, Coding, and Symbol Search) and 5 supplemental subtests (Information, Word Reasoning, Picture Completion, Arithmetic and Cancellation). The WISC-IV allows to have one FSIQ and following four Indexes: Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index and Processing Speed Index.

2.3. Statistical analyses

All analyses in the present study are consistent with the steps in testing for string factorial invariance; all procedures were based on the analysis of MACS within the framework of CFA (Confirmatory Factor Analysis) modelling: The EQS 6.1 (Bentler, 2005) program were used for all analyses.

In particular, to test equivalence of WISC-IV construct across two sex needed encompassed a series of hierarchically ordered steps according the procedure of Chen, Sousa, and West (2005); tests of measurement invariance examine whether the same construct has been measured across different two gender groups.

We followed testing increasingly stringent levels of constrained equivalence across the groups. The most basic level of measurement invariance is *configural invariance* (model 1): the same item must be an indicator of the same latent factor in each group, however the loadings-first and second-order factors — and intercepts can differ across the groups. This means latent variables are present in the examined groups. The second level of invariance is factor *loadings invariance* that represents the strength of the linear relation between each factor and its associated items (first-order factor model: model 2) and between first- and second-order factors (second-order factor loadings: model 3). When the loadings are equal across groups, means that the unit of measurement is identical. However, the factor means still cannot be compared across groups, because the factors have not a common measure origin.

The third level of invariance that represent the origin of the factor is *intercept invariance*, about measured variables (model 4) and first-order latent factors (model 5). The four level concerns the invariance of residual variances of first-order factors in which factor loadings, intercepts of measured variables, intercepts of the first-order latent factors were constraints to be equal across groups (model 6). The last level concerns the invariance of residual variance of observed variables in which factor loadings, intercepts of measured variables, intercepts of the first-order latent factors, residual variances of first-order factors and of the measured variables were constrained to be equal across groups (model 7).

Each level had more constraints than the previous one. Seven multigroup models were tested across males and females, each representing an increasingly more restricted parameterization than its predecessor, so these models are said to be hierarchically nested.

The evaluation of model fit was based on followed multiple indices of model fit: we preferred the *Yuan–Bentler Scaled Statistic* (Y-Bχ^2), rather than the uncorrected Chi-squared (χ^2) statistic because our data are non-normally distributed; the *Root Mean Square Error of Approximation* (RMSEA; Steiger, 1990), is an index of approximate fit that tells us how well the model fit the populations covariance matrix: a value of less than .05 indicates good fit, equal to .0 indicates exact fit, a value ranging from .05 to .08 indicates a mediocre fit, while a value greater than .08 indicates no fit. It is also possible to add a 90% confidence interval to the RMSEA, in which case the lower limit of the interval should be lower than .05 while the upper limit should be lower than .08 to represent a good fit; the *Standardized Root Mean Square residual* (SRMR; Joreskog & Sorbom, 1984; Kline, 2005), is an index of the average discrepancy among the residuals of the observed and fitted covariance matrices: a good model should have a SRMR less than .05 (Byrne, 1998), however values as high as .08 are deemed acceptable (Hu & Bentler, 1999); the *Comparative Fit Index* (CFI; Bentler, 1990) ranges in values from 0 to 1.00, with a value of .95 serving as the rule-of-thumb cut-point of acceptable fit (Hu & Bentler, 1999).

When models are nested, they can be compared in pairs computing the difference in their overall Chi-square (ΔSBχ^2) values and the related degrees of freedom: if this value is statistically significant, it suggests that the constraints specified in the more restrictive model do not hold. However, there are researches (e.g., Cheung & Rensvold, 2002) have argued that this Delta Chi-square value "is as a sensitive to sample size and nonnormality as the chi-square statistic itself, thereby rendering it an impractical and unrealistic criterion on which to base evidence of invariance" (Byrne & Stewart, 2006, pp. 290).

Based in the examination of properties related to 20 goodness-of-fit indices within the context of invariance testing, Cheung and Rensvold (2002) recommended that the *CFI Difference Test* (or ΔCFI) provides the best information in determining evidence of measurement invariance and they suggested that its difference value should not exceed .01. Again, Meade, Johnson, and Braddy (2006) proposed to use the difference between *Mac Donald's* non centrality indexes (ΔMcDonald) to evaluate measurement invariance. A McDonald's fit index is a fit Index based on covariance matrix and means and a value ΔMcDonald smaller than or equal to .02 (in absolute terms) indicates that the null hypothesis of invariance should be rejected (Cheung & Rensvold, 2002; Meade et al., 2006).

In the third step, we conducted *tests for latent mean differences*. This step is an important consequence of invariance findings that it enabled us to subsequently test for differences in the latent factor means with respect to both the lower and higher order factors. In estimating latent mean differences on the Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed, all first-order factor loadings and observed variable intercepts were constrained equal across groups. Similarly, in estimating the latent mean difference on General Intelligence, all first- and second-order factor loadings, as well as observed and latent factor intercepts, were constrained equal across two groups of males and females.

To test the latent construct mean differences, a combined mean and covariance structure model was been used (Bentler, 1990). To estimate the difference between the factor means, one group (the females of this sample) was been chosen as a reference or baseline group and its latent means are set to zero. So, the latent means of the other group (males), which actually represent the difference between the factor means in the two groups, are estimated. The significance test (z test) provides a test for significance of the difference between the means of the two groups on the latent construct.

Finally, to test observed means differences five ANOVAs one-way on ten core subtests, four indexes and Full Scale IQ were computed. The