Personality and Individual Differences 53 (2012) 126–131

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Sex differences in g: An analysis of the US standardization sample of the WAIS-III

Paul Irwing *

Manchester Business School East, The University of Manchester, Booth Street West, Manchester M15 6PB, United Kingdom

ARTICLE INFO

Article history:
Available online 31 May 2011

Keywords:
Sex differences
Multi-group confirmatory factor analysis
Measurement invariance
Intelligence
WAIS

ABSTRACT

This study employed both hierarchical and Bi-factor multi-group confirmatory factor analysis with mean structures (MGCFA) to investigate the question of whether sex differences are present in the US standardization sample of the WAIS-III. The data consisted of age scaled scores from 2450 individuals aged from 16 to 89 years. The findings were more or less uniform across both analyses, showing a sex difference favoring men in *g* (0.19–0.22*d*), Information (0.40*d*), Arithmetic (0.37–0.39*d*) and Symbol Search (0.40–0.30*d*), and a sex difference favoring women in Processing Speed (0.72–1.30*d*).

1. Introduction

The question of whether there is a sex difference in general cognitive ability is a matter of considerable controversy. Richard Lynn has made three important contributions to debate on this issue. Firstly, he has proposed that there is a male advantage on *g* in adults of about 3–5 IQ points (Lynn, 1994, 1999), secondly that there is a developmental trend whereby, while among children up to the age of 16 years the sex difference in overall intelligence is negligible, the male advantage begins to appear at the age of 16 and increases into early adulthood. For convenience, we will dub this the developmental theory of sex differences in cognitive ability. Thirdly, he has questioned the overwhelming consensus that there is greater male variability (Irwing & Lynn, 2005; Johnson, Carothers, & Deary, 2008). This paper will test all three of these propositions in the US standardization sample of the WAIS-III.

From discussions of the issue you might think that the evidence is overwhelmingly against the developmental theory of sex differences (e.g. Ceci, Williams, & Barnett, 2009). In fact, a simple examination of empirical findings shows that by far the majority of the evidence favors a mean male advantage in adulthood and that its emergence follows a developmental trend (e.g. Irwing & Lynn, 2005; Jackson & Rushton, 2006; Johnson & Bouchard, 2007; Lynn, 1994, 1999; Lynn & Irwing, 2004). There are studies which apparently support a null sex difference, or even a female advantage among adults, though most of these studies have used multi-group confirmatory factor analysis (MGCFA) (e.g. Dolan et al., 2006; Keith, Reynolds, Patel, & Ridley, 2008; van der Sluis et al., 2006).

The confused state of debate on this issue is perhaps attributable to a number of methodological problems, which any study of sex differences needs to address. Firstly, there is a problem of selection biases which may mean that any given sample is not equally representative of males and females (Madyastha, Hunt, Deary, Gale, & Dykiert, 2009). Secondly, findings are method dependent, and there are strong arguments favouring MGCFA as the preferred form of analysis (Dolan et al., 2006). In particular, a number of criticisms of the method of correlated vectors have been made (e.g. Ashton & Lee, 2005; Lubke, Dolan, & Kelderman, 2001), such that conclusions depending on this method must be regarded as suspect. Thirdly, there is the issue of the quality of tests and exactly what they measure. Fourthly, the establishment of measurement invariance and lack of bias represent prerequisites for the unequivocal demonstration of sex differences (Meredith, 1993). Fifthly, there is strong evidence that g is not normally distributed (Johnson et al., 2008). Unfortunately, no study, including the current one is immune from all these difficulties.

It was probably Gustafsson who first suggested that MGCFA should be the preferred method of analyzing group differences in intelligence. Which method is appropriate is dependent on which model of intelligence is veridical. Certainly MGCFA is compatible with the consensus hierarchical factor models of human cognitive abilities. Apart from compatibility, MGCFA has many other advantages over alternatives such as the method of correlated vectors or exploratory factor analysis (Bollen, 1989). It may, therefore, seem damaging that studies using MGCFA have uniformly failed to support a mean male advantage in *g*. However, there are a number of complications in conducting such analyses. It has been shown by Molenaar, Dolan, and Wicherts (2009) that large samples are

* Tel.: +44 161 8320604.
E-mail address: paul.irwing@mbs.ac.uk

0191-8869/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2011.05.001

required to attain sufficient power in order to detect a mean difference in MGCFA models. Here, we have such a large sample, and in order to ensure sufficient power we carry out the analysis in the entire sample aged 23 years and older. A more profound difficulty is that most analyses have failed to separate out measurement issues from structural analyses. In doing so, authors have simply followed recommended practice (Chen, Sousa, & West, 2005). The problem is that for cross group comparisons to be valid scalar invariance must hold (Widaman & Reise, 1997). To establish scalar invariance multiple congeneric measures at the first order factor level are required (Widaman & Reise, 1997), but to date, no study including the current one, has had access to multiple measures. However, we adopt a somewhat novel solution by simply recognizing that testing of metric invariance is the most that we can achieve with only one measure for each construct.

Probably the most serious problem in validly testing for mean differences in MGCFA models is that factors are correlated, and therefore order of testing influences the conclusion. The problem is closely analogous to that presented by post hoc testing in multivariate analysis of variance. Here, in order to achieve an unambiguous conclusion, we present two solutions to this problem. The first followed the practice in stepdown analysis of prioritizing the order of testing according to a mixture of theoretical and practical criteria. We then used a Bonferroni correction in order to control for type 1 error. In the second, we used a Bi-factor model which removes the problem of correlated factors by orthogonalizing them.

In short we use one of the best samples, the doyen of psychometric tests of general cognitive ability, and a novel testing procedure in order to examine Lynn's developmental theory.

2. Method

2.1. Sample

The sample analyzed in this study is the American standardization sample of the WAIS-III[1]. This consists of 2450 individuals aged from 16 to 89 years. The data consist of sex differences in age scaled scores provided by the Psychological Corporation. The standardization sample was designed to be representative of the US population according to the 1995 census, with regard to age, sex, ethnicity, educational level and geographic region (US Bureau of Census, 1995). Three categories of adults were excluded from the sample: individuals with sensory or motor deficits that might compromise the validity of test scores; individuals fitting criteria for drug or alcohol dependency or who were on medication; and individuals with known or possible neuropsychological disorders. These exclusions would not seem to impair the suitability of the sample for the analysis of sex differences.

2.2. Measures

The WAIS-III contains 13 subtests and a Full Scale IQ, a Verbal IQ and a Performance IQ, like its predecessors. It also provides measures of four factors: Verbal Comprehension (Vocabulary, Similarities, Information, Comprehension), Perceptual Organization (Picture Completion, Block Design, Matrix Reasoning, Picture Arrangement), Working Memory (Arithmetic, Digit Span, Letter-Number Sequencing), and Processing Speed (Digit Symbol – Coding, Symbol Search). Object Assembly is an optional test, but the current analysis placed it on the Perceptual Organization factor, in common with some other analyses (Dolan et al., 2006). Average split-half reliability coefficients across the 13 age groups were .98 for Full Scale IQ, .97 for Verbal IQ and .94 for Performance IQ. The average reliabilities for the individual subtests ranged from .93 (Vocabulary) to .70 (Object Assembly).

Descriptive statistics for sex differences in the American WAIS III data are given in Table 1, which shows the means, standard deviations, and sample sizes for male and female subtest and scale scores on the WAIS-III, together with Cohen's d (the male mean score minus the female mean score divided by the within-group standard deviation). Multivariate ANOVA revealed main effects of sex for both the subtests ($F(14, 1284) = 30.38$, $p < .001$) and scale scores ($F(4,1294) = 46.70$, $p < .001$). Twelve of the 14 subtest difference scores are in favor of males (six significant at the .001 level), and two are in favor of females (both significant at the .01 level). Cohen's d for the Full-Scale IQ score is .185 in favor of males.

3. Results

We have analyzed the data using two different models for reasons explained above. Because, in 1151 cases, there were missing data for Letter–Number Series, we used Full Information Maximum Likelihood estimation for all analyses, which broadly conforms with best practice (Schafer & Graham, 2002). In all cases, we test for measurement invariance in the order: (1) configural invariance; and (2) metric invariance (for the reasons given above, we do not consider tests for scalar invariance to be logical as applied to this data set). As a third step, we constrained all mean and intercept differences across sex to zero and then, in subsequent models, allowed for mean differences based on both theory (Bollen, 1989) and modification indices (Jöreskog & Sorbom, 2001). Finally, in the Bi-factor model we tested for sex differences in factor variances. The theory and logic of testing for measurement invariance is extensively detailed elsewhere (e.g. Meredith, 1993; Widaman & Reise, 1997) so we do not repeat this here.

There is no fully satisfactory answer to the question of model fit, particularly as this applies to testing for measurement invariance (Yuan, 2005). Moreover, with Full Information Maximum Likelihood, the only available fit indices are the likelihood ratio statistic and the root mean square error of approximation (RMSEA). We rely partly on the simulations of Hu and Bentler (1998, 1999), which suggest that in order to assess absolute fit, a cut-off point of about .06 is appropriate for the RMSEA. Decline in model fit at a given stage of the invariance analysis indicates that the assumptions of invariance do not hold in the constrained parameters (French & Finch, 2006). To assess possible decline in model fit, we rely on the conclusion of Cheung and Rensvold (2002). Their primary recommendation is that changes of equal to or less than −0.01 for CFI indicate that invariance holds. However, since this statistic is not available, we suggest a comparable cut-off value of 0.013 for the RMSEA, based on their findings. Though conventionally the χ^2 difference statistic has been proposed as a measure of decrease in fit between nested models, it too has been demonstrated to be sensitive to sample size (Kelloway, 1995), and therefore it has been argued to be inferior to other metrics for comparison of nested models (Cheung & Rensvold, 2002).

3.1. Hierarchical MGCFA

First we consider results for the hierarchical MGCFA factor model shown in Fig. 1. We analyzed this using the subsample aged 23 years or older ($N_{male} = 902$, $N_{female=}$ 1053), since, according to data presented in Lynn and Irwing (2004), together with developmental studies of brain tissue, we surmise that this is the age at which sex differences probably attain their full adult value. All invariance analyses considered parameters in the first- and

[1] Standardization data from the *Wechsler Adult Intelligence Scales*® – *Third Edition.*