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*.

Table 1
Univariate means, standard deviations and Cohen's *d*.

Scale/subtest	Males			Females			*d*	*F*
	N	M	SD	N	M	SD		
Full-Scale IQ	603	134.41	25.83	696	129.42	27.92	.185	<.001
Verbal Comprehension	1147	41.44	10.59	1303	38.96	10.65	.233	<.001
Vocabulary	1147	10.08	2.98	1303	9.96	3.04	.038	.230
Similarities	1147	10.19	3.00	1303	9.91	3.00	.095	.055
Information	1147	10.67	3.04	1303	9.40	2.82	.433	<.001
Comprehension	1147	10.50	2.91	1303	9.69	2.98	.276	<.001
Perceptual Organization	1147	41.26	9.71	1303	39.19	9.25	.219	<.001
Picture Completion	1147	10.15	3.05	1303	9.92	2.97	.076	.077
Block Design	1147	10.51	3.10	1303	9.70	2.78	.274	<.001
Matrix Reasoning	1147	10.23	2.95	1303	9.85	2.95	.130	.003
Picture Arrangement	1147	10.38	3.13	1303	9.72	2.94	.217	<.001
Working Memory	603	31.27	7.44	696	29.51	7.41	.238	<.001
Arithmetic	1147	10.68	3.21	1303	9.45	2.96	.399	<.001
Digit Span	1147	10.15	3.03	1303	9.95	3.04	.069	.095
Letter–Number	603	10.18	3.14	696	9.92	3.08	.083	.135
Processing Speed	1147	19.15	5.21	1303	20.82	5.54	−.308	<.001
Digit Symbol	1147	9.27	2.81	1303	10.60	3.04	−.456	<.001
Symbol Search	1147	9.89	2.94	1303	10.21	3.05	−.108	.006
Object Assembly	1147	10.09	3.02	1303	9.96	3.05	.040	.391

The last column contains *p* values based on the multivariate *F*-test, *d* represents the males' minus the females' means scores divided by the within-groups standard deviation.

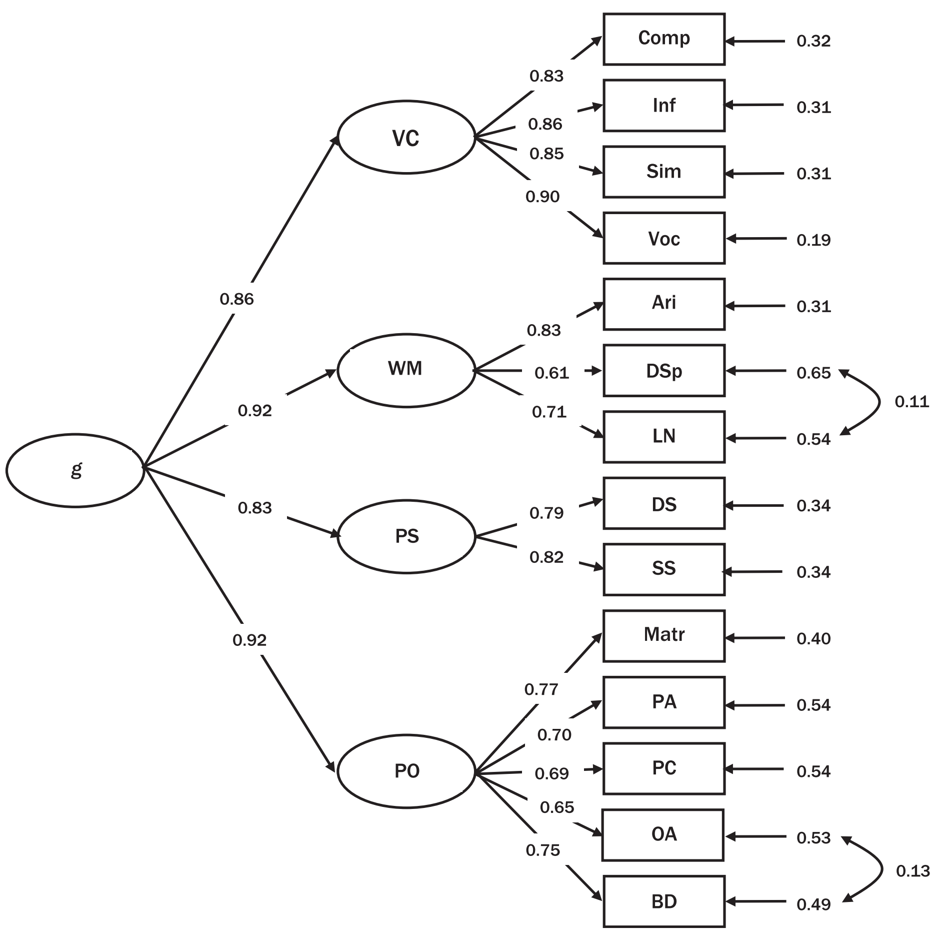

Fig. 1. The WAIS-III second-order confirmatory factor model with mean structures (M_8) – common metric completely standardized solution (VC, Verbal Comprehension; WM, Working Memory; PS, Processing Speed; PO, Perceptual Organization; Comp, comprehension; Inf, information; Sim, similarities; Voc, Vocabulary; Ari, Arithmetic; DSp, Digit Span; LN, Letter–Number sequencing; DS, Digit Symbol; SS, Symbol Search; Matr, Matrix Reasoning; PA, Picture Arrangement; PC, Picture Completion; OA, Object Assembly; BD, Block Design).

second-order factor models simultaneously. For the configurally invariant model (same factor pattern), the RMSEA was within the specified cut-off, clearly demonstrating that the model provides a good fit to the data (χ^2 = 569.1, *df* = 134; *p* < .001; RMSEA = .055). For the fully metrically invariant model the fit effectively improved as indicated by a reduction in the RMSEA (χ^2 = 597.2, *df* = 147;

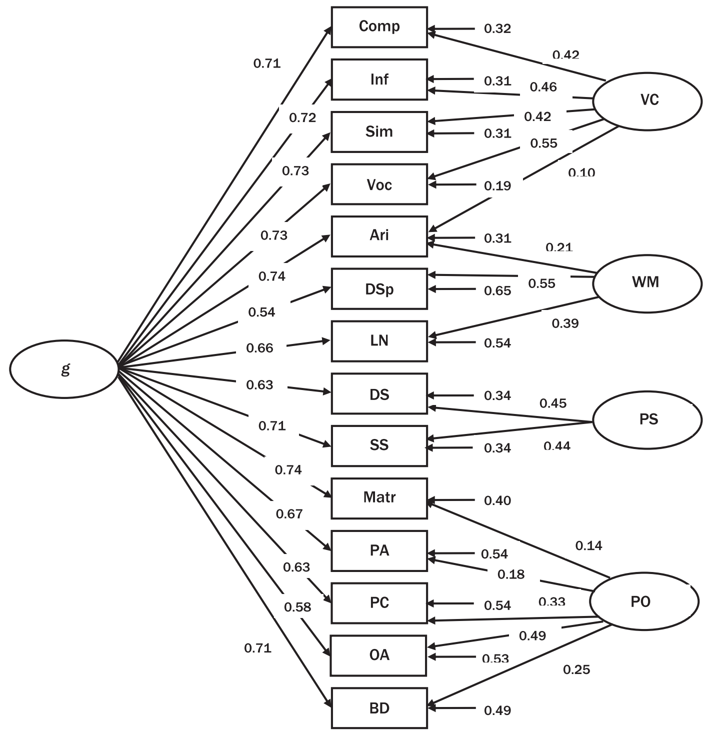

Fig. 2. The WAIS-III Bi-factor model with mean structure (M_{16}) = common metric completely standardized solution (abbreviations as for Fig. 1).

$p < .001$; RMSEA = .054, ΔRMSEA = −.001), so we conclude that metric invariance is convincingly demonstrated.

At the next step, we constrained all subtest means (intercepts) and factor means to equality in both males and females. This resulted in a dramatic decrement in fit ($\chi^2 = 1254.6$, $df = 171$; $p < .001$; RMSEA = .081, $\Delta\chi^2 = 657.4$, $df = 24$, $p < .001$). This clearly establishes that there are mean differences in either factor or scale scores across sex.

There are well established sex differences in Processing Speed, Information and Arithmetic, while *g* is the focus of the investigation (e.g. Hedges & Nowell, 1995; Held, Alderton, Foley, & Segal, 1993; Lynn, Irwing, & Cammock, 2002; Majeres, 2007). Following the logic of stepdown analysis the means (intercepts) for each of these variables was released first. Next, an inspection of modification indices (MI) suggested that the intercept for Symbol Search (MI = 77.4) should be released. This model provided equivalent fit to that of the metrically invariant model in terms of the RMSEA index ($\chi^2 = 709.4$, $df = 166$; $p < .001$; RMSEA = .058, ΔRMSEA = 0.004), so we accepted this model. Next we released the intercepts for Verbal Comprehension, Working Memory and Perceptual Organization which lead to a non-significant change in the likelihood ratio ($\Delta\chi^2 = 2.09$, $df = 3$, $p > .05$), so we may conclude that g adequately explains the residual mean differences for these second-order factors.

The differences in the means expressed in *d* scores are 0.19 ($t = 3.88$, $p = .00005$), −0.72 ($t = 15.54$, $p < .00001$), 0.40 ($t = 7.66$, $p < .00001$), 0.37 ($t = 14.14$, $p < .00001$), and 0.40 ($t = 32.12$, $p < .00001$) for *g*, Processing Speed, Information, Arithmetic and Symbol Search, respectively (a negative score denotes a female advantage). The Bonferroni corrected probability is .002 in order to maintain the probability of a type 1 error at .01, so clearly all the *d*-score differences are significant.

3.2. Bi-factor models

Although the hierarchical factor model corresponds well to some conceptualizations of the structure of intelligence, it has some disadvantages. Because all the factors are correlated, this can lead to ambiguities in the interpretation of such models. In Bi-factor models, the factors are uncorrelated, which in principle greatly simplifies interpretation. For these reasons, Carroll (2003), for example, has favoured such models. Therefore, using the same strategy and logic of analysis as presented above, we examined Bi-factor models (see Fig. 2). With regard to all tests of measurement invariance, the Bi-factor model supported identical conclusions to those derived from the hierarchical factor model. The mean score sex differences were highly similar at 0.22, .0.40, .0.39 and 0.30 for *g*, Information, Arithmetic and Symbol Search, respectively, but the estimated difference for Processing Speed, at 1.30, was substantially larger.

The Bi-factor model has another advantage in that it greatly simplifies testing for equality of factor variances. We applied equality constraints to variances and error variances in the metrically invariant model. All means and intercepts were also constrained except for those five which were significantly different ($\chi^2 = 629.3$, $df = 176$; $p < .001$; RMSEA = .051, $\Delta\chi^2 = 71.8$, $df = 25$, $p < .001$). We then sequentially released constraints on each of the variances one at a time. The variance ratios and associated chi-square differences were: *g* (VR = 1.04, $\Delta\chi^2 = 0.32$, $df = 1$, $p = 0.572$), Verbal Comprehension (VR = 1.03, $\Delta\chi^2 = 0.07$, $df = 1$, $p = 0.791$), Working Memory (VR = 1.39, $\Delta\chi^2 = 5.07$, $df = 1$, $p = 0.024$), Processing Speed (VR = 0.65 ,$\Delta\chi^2 = 9.07$, $df = 1$, $p = 0.003$), and Perceptual Organization (VR = 1.14, $\Delta\chi^2 = 0.45$, $df = 1$, $p = 0.502$). We can thus conclude that there are no significant differences in variability between males and females on g, Verbal Comprehension, and Perceptual Organization, while there is significantly greater male variability on Working Memory at the .05 level, and significantly greater variability in females on Processing Speed at the .01 level.

3.3. Sex differences across age in g

There is evidence that education has an effect on intelligence (Dolan et al., 2006; Johnson, Deary, & Iacono, 2009), so it could be argued that the observed sex difference in *g* favouring males may be attributable to the older age groups in the sample in which women would have been exposed to less education than males.

In order to test for sex differences in *g* across age, we first calculated a composite *g* score using factor score regression. This approach could be criticized in that it is known that composite measures of *g* are potentially contaminated with non-*g* variance. However, provided our results are not greatly discrepant from those obtained from the latent variable analyses, we can be confident that the parameter estimates observed with the composite score are in this instance closely equivalent. The sample was divided into 13 age bands with 200 participants in all age bands except the oldest two which comprised 150 and 100, respectively. We carried out an analysis of variance with the g composite score as the dependent variable, and sex and age group as the independent variables. There was a significant mean difference for sex ($d = 0.177$, $F = 17.95$, $df = 1$, $p < 0.001$), but neither the age nor the interaction term were significant. To provide a direct test of the effects of exposure to education we next controlled for length of education divided into five levels from ⩽8 to ⩾16 years. The effect of sex remained significant ($F = 14.41$, $df = 1$, $p = .001$), and Cohen's d (0.154) reduced only marginally. Consequently, the argument that the sex difference in *g* is attributable to differential experience of education does not appear to hold.

Figure 3 shows a plot of the data which shows some interesting features of the profile of *g* across age. Firstly, although not significant, there is a trend whereby the sex difference in *g* increases across age from 17–23 years, as predicted by the developmental theory of sex differences. Secondly, across age from 23–60 years, male *g*-scores appear to follow a V shaped trend, while over the same period female *g*-scores follow an inverted V.

4. Discussion

The MGCFA and Bi-factor analyses both show the existence of a sex difference favoring men in *g*, Information, Arithmetic, and Symbol Search and a sex difference favoring women in Processing Speed. The sex difference effect sizes were highly similar except that the estimate for Processing Speed was substantially larger in the Bi-factor model.

Our results that the females have an advantage on Processing Speed, while males have an advantage on Information and Arithmetic replicate the findings of a number of other studies (e.g. Hedges & Nowell, 1995; Held et al., 1993; Lynn et al., 2002; Majeres, 2007). The large female advantage of 0.72–1.30*d* on Processing Speed is particularly notable. The magnitude of this effect arises partially because *g* is not masking sex differences in this analysis (Johnson & Bouchard, 2007). Nevertheless, this finding does support the argument of Majeres (2007) that because the female brain is highly specialized for processing phonologically coded information, this provides a female advantage on a range of cognitive tasks including perceptual speed, digit-symbol substitution, numerical computation, spelling ability and word-level reading. Neuroimaging studies also show that during phonological processing women evidence greater right hemisphere activation (e.g. Pugh et al., 1997). If women devote more right

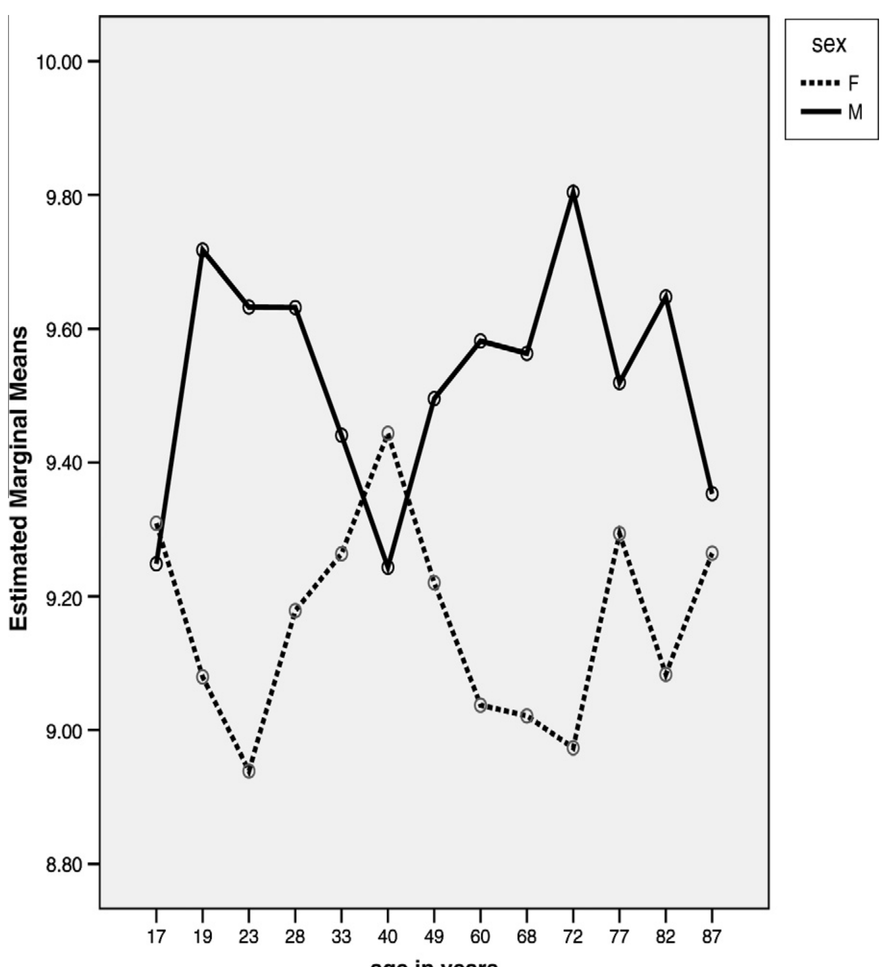

Fig. 3. Age profile of *g*-scores by sex.

hemisphere brain tissue to phonological processing, and men devote more to visual rotation, then this might explain the trade-off observed by Johnson and Bouchard (2007) on what they refer to as a rotation-verbal dimension.

Contrary to some previous findings (e.g. Hedges & Nowell, 1995; Johnson et al., 2008), we did not find greater variability in male scores on g, Verbal Comprehension, or Perceptual Organization, and while we did find greater male variability on Working Memory, there was even greater female variability on Processing Speed. Our results add to the numerous inconsistencies in findings on sex differences in variability. There are some possible reasons for these discrepancies. Firstly, no latent variable analysis of the issue has found greater male variability. This may be attributable to known difficulties with composite variables. For example, during development up to about 14 years of age, males score lower on Verbal ability, and higher on Visuo-Spatial ability than do females. This in itself would lead to greater apparent variability in males on a composite containing these factors. Equally, if there are greater differences in developmental lags among males, this would also produce greater male variability until they enter adulthood. Alternatively, the findings may be due to a lack of power.

Although the WAIS-III is a very highly regarded cognitive battery, nevertheless it does suffer some limitations for the estimation of sex differences. In particular, 3-D mental rotation, for which there is a large male advantage (e.g. Voyer, Voyer, & Bryden, 1995), is not tested in the WAIS-III, and therefore the WAIS-III is likely to provide an underestimate of the sex difference in *g*. We also found that between the ages of 23 and 60 that the sex difference was strongly attenuated (see Fig. 3). One interpretation would be that successful males in these age ranges are harder to contact than females, while intelligent females are more likely to volunteer. Despite rigorous attempts at random sampling it would not be surprising if the WAIS-III standardization sample was subject to such selection effects, which would lead to an underestimate of the sex difference in g.

The WAIS-III manual also documents that extensive procedures were used in the construction of the test in order to eliminate gender bias. The methods of expert opinion and differential item functioning (DIF) were both used for this purpose. It is impossible to know exactly how these procedures were employed. However, it is well established that expert opinion is not a good basis to establish bias in items (Smith & Smith, 2005). DIF analyses will also tend to remove unbiased items unless item pools are truly unidimensional and this is something that is hard to establish (Embretson & Reise, 2000). Taking all these considerations together, there are some grounds to think that the WAIS-III, despite its excellent psychometric properties, may underestimate sex differences in cognitive abilities.

In conclusion, our findings provide further support for Lynn's developmental theory of sex differences, and suggest that the consensus view that there is greater male variability in cognitive abilities requires further investigation.

References

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence, 33*, 431–444.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The Scientific Study of General Intelligence*. Amsterdam: Elsevier.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's Underrepresentation in Science. Sociocultural and Biological Considerations. *Psychological Bulletin, 135*, 218–261.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*, 471–492.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 235–255.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & van der Sluis, S. (2006). Multi-group covariance and mean structure modelling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence, 34*, 193–210.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13*, 378–402.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41–45.

Held, J. D., Alderton, D. L., Foley, P. P., & Segal, D. O. (1993). Arithmetic reasoning gender differences: Explanation found in the Armed Forces Vocational Aptitude Battery (ASVB). *Learning and Individual Differences, 5*, 171–186.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structural modelling: Sensitivity to underparametrized model misspecification. *Psychological Methods, 3*, 424–453.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure modelling: Conventional criteria versus new alternatives. *Structural Equation Modelling, 6*, 1–55.

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology, 96*, 505–524.

Jackson, D. N., & Rushton, J. P. (2006). Males have greater *g*: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence, 34*, 479–486.

Johnson, W., & Bouchard, T. J. Jr., (2007). Sex differences in mental abilities: *g* Masks the dimensions on which they lie. *Intelligence, 35*, 23–39.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3*, 518–531.

Johnson, W., Deary, I., & Iacono, W. J. (2009). Genetic and environmental transactions underlying educational attainment. *Intelligence, 37*, 466–478.

Jöreskog, K. G., & Sorbom, D. (2001). *LISREL 8: User's reference guide*. Lincolnwood, ILL: Scientific Software International Inc.

Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock Johnson III tests of cognitive abilities. *Intelligence, 36*, 502–525.

Kelloway, K. E. (1995). Structural Equation Modelling in Perspective. *Journal of Organizational Behavior, 16*, 215–224.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research, 36*, 299–324.

Lynn, R. (1994). Sex differences in brain size and intelligence: A paradox resolved. *Personality and Individual Differences, 17*, 257–271.

Lynn, R. (1999). Sex differences in intelligence and brain size: A developmental theory. *Intelligence, 27*, 1–12.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence, 32*, 481–498.

Lynn, R., Irwing, P., & Cammock, T. (2002). Sex differences in general knowledge. *Intelligence, 30*, 27–39.

Madyastha, T. M., Hunt, E., Deary, I. J., Gale, C. R., & Dykiert, D. (2009). Recruitment modeling applied to longitudinal studies of group differences in intelligence. *Intelligence, 37*, 422–427.

Majeres, R. L. (2007). Sex differences in phonological coding: Alphabet transformation speed. *Intelligence, 35*, 335–346.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.

Molenaar, D., Dolan, C. V., & Wicherts, J. M. (2009). The power to detect sex differences in IQ scores using Multi-Group Covariance and Means Structure Analysis. *Intelligence, 37*, 396–404.

Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Fletcher, J., et al. (1997). Predicting reading performance from neuroimaging profiles: The cerebral basis of phonological effects in printed word identification. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 145–151.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Smith, J. M., & Smith, P. (2005). *Testing people at work: Competencies in psychometric testing*. London: Blackwell.

US Bureau of Census (1995). *Current population survey*, March, 1995 (Machine readable data file). Washington, DC: US Bureau of Census (Producer/ Distributor).

Van der Sluis, S., Posthuma, D., Dolan, D. V., de Geus, E. J. C., Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence, 34*, 273–290.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*, 250–270.

Widaman, K. F., & Reise, K. F. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant & M. Windle (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*, 115–148.