Available online at www.sciencedirect.com

Intelligence 34 (2006) 273–289

Sex differences on the Dutch WAIS-III

Sophie van der Sluis [a,*], Danielle Posthuma [b], Conor V. Dolan [a], Eco J.C. de Geus [b], Roberto Colom [c], Dorret I. Boomsma [b]

[a] *Department of Psychology, FMG, University of Amsterdam, Roeterstraat 15, 1018 WB Amsterdam, The Netherlands*

[b] *Department of Biological Psychology, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands*

[c] *Facultad de Psicologia, Universidad Autonoma de Madrid, 29049 Madrid, Spain*

Received 31 May 2005; received in revised form 14 August 2005; accepted 20 August 2005
Available online 6 October 2005

Abstract

Using multi-group covariance and means structure analysis (MG-CMSA), this study investigated whether sex differences were present on the Dutch WAIS-III, and if so, whether these sex differences were attributable to differences in general intelligence (g). The sample consisted of 294 females and 228 males between 18 and 46 years old. Both first and second order common factor models were fitted, the latter including g as second order factor. The results indicated that on the level of the subtests, females outperform males on Digit–Symbol Substitution, and males outperform females on Information and Arithmetic. In addition, the subtests Information proved to be biased in favor of males. With respect to the first order common factors, no sex differences were found with respect to the factor Verbal Comprehension (once Information was effectively removed from the model). Yet, males outperformed females on the factors Working Memory and Perceptual Organization, and females outperformed males on Perceptual Speed. These sex difference on the level of the first order common factors were however not attributable to sex differences in g. Summarizing, the present study showed that males and females do differ with respect to specific cognitive abilities, but that g cannot be viewed as the source of these differences.

Keywords: Sex differences; Multi-group covariance and means structure analysis; Measurement invariance

1. Introduction

In comparing groups with respect to their cognitive ability, one question of interest is whether mean group differences observed on specific cognitive subtests are attributable to differences in general intelligence, or 'g'. g is defined as the general factor underlying all aspects of cognitive ability. It is assumed that all tests of cognitive ability measure g to some extent (Carroll, 1997; Jensen, 1981, 1998). Yet, an overall mean difference between groups on a collection of IQ (sub)tests, even if significant, is not necessarily attributable to differences in g (Jensen, 1998). Mean group differences may also be due to bias at the level of the subtests, or indicate sex differences in specific cognitive abilities (i.e., broad primary factors of intelligence such as spatial ability, or verbal ability), rather than sex differences in general mental ability. The present study is concerned with sex differences in cognitive ability, and especially with the role of g in these differences.

Several studies report mean differences between males and females with respect to specific cognitive

[*] Corresponding author. Biological Psychology, VU Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.

E-mail addresses: S.van.der.Sluis@psy.vu.nl (S. van der Sluis), C.V.Dolan@uva.nl (C.V. Dolan).

0160-2896/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2005.08.002

abilities. For example, females have been reported to outperform males on perceptual speed (Kimura, 1999) and verbal ability (although see Hyde & Linn, 1988), while males outperform females on spatial ability (Voyer, Voyer, & Bryden, 1995), and on tests measuring general knowledge, such as the Information subtest of the Wechsler intelligence batteries (e.g., Born & Lynn, 1994; Jensen & Reynolds, 1983; Lynn, 1998). These differences on specific abilities notwithstanding, the received view is that males and females do not differ with respect to general intelligence or *g*. In the last decade, however, Rushton (1992), Lynn (1994, 1999), and Nyborg (2003) challenged this view and stated that males show on average a 4 IQ-point advantage over females. Lynn argued that these differences have long been underestimated, because researchers studying sex differences in adolescents and young adults, have failed to take into account the fact that females on average mature somewhat earlier than males. According to Lynn, males do show slight but consistent cognitive advantage over females once males and females have reached adulthood, and the maturational advantage of females has disappeared. Colom and Lynn (2004) recently offered support for this differential maturation hypothesis.

Sex differences in psychometric IQ have been studied in adolescent and adult samples using the method of correlated vectors. Jensen (1998) designed this method to investigate the role of *g* in the relationship between subtest scores on the one hand, and variables such as sex, race, and scholastic attainment on the other. Within this procedure, the *g*-loadings of every subtest are estimated in the groups separately either in a Schmid–Leiman factor analysis (Schmid & Leiman, 1957), or through the extraction of the first unrotated principal factor. Subsequently, the correlation is calculated between these vectors of *g*-loadings across the groups. Large correlations have been reported between the vectors of *g*-loadings of males and females (.99, Aluja-Fabregat, Colom, Abad, & Juan-Espinosa, 2000; .97, Carretta & Ree, 1995; .99, Carretta & Ree, 1997; .99, Colom, Juan-Espinosa, Abad, & García, 2000; .95, Escorial, Juan-Espinosa, García, & Rebollo, 2003). These large correlations have been taken to mean that the *g*-factor is (virtually) identical across sex. However, as a test of measurement invariance (see below), this is insufficient.

In addition, researchers have investigated whether tests with higher *g*-loadings show greater sex differences by calculating the correlation between the vector of *g*-loadings as estimated across sex, and the vector of standardized mean differences between males and females. These correlations are usually low (e.g., .00, Colom et al., 2000; .06, Colom, Abad, García, & Juan-Espinosa, 2002; .116, Jensen, 1998), as are the correlations between *g* and sex (e.g., −.25 to .13, Aluja-Fabregat et al., 2000; −.07 to .36, and −.09 to .35, Colom et al., 2000). These results have led researchers to conclude that *g* is not the source of the observed differences between males and females on the level of the subtests.

Another method that is used to investigate group differences in psychometric IQ is multi-group covariance and means structure analysis (MG-CMSA). In MG-CMSA, one first fits the confirmatory factor model separately in the different groups. The specific structure of this baseline model is based on empirical or theoretical considerations. Subsequently, different hypotheses can be evaluated and compared by constraining parameters to be equal across groups. This method provides a comprehensive, model-based means to investigate whether the general factor *g* is the only, or main, source of group differences, or whether the primary (first-order) factors are the main source of group differences. Besides, MG-CMSA provides the possibility to detect bias at the level of the subtests. MG-CMSA has been applied in the investigation of ethnic differences in psychometric IQ (e.g., Dolan & Hamaker, 2001; Dolan, Roorda, & Wicherts, 2004; Gustafsson, 1992; Lubke, Dolan, Kelderman, & Mellenbergh, 2003), and in the study of the Flynn effect (Wicherts, et al., 2004). Recently, Dolan, Colom, Abad, Wicherts, Hessen, and van der Sluis (submitted for publication) used MG-CMSA to study sex differences on Spanish WAIS-III data obtained from 16–35 year old males and females. They found that males outperformed females on five of the fourteen subtests (Arithmetic, Digit Span, Information, Letter Number Sequencing, and Block Design), and that four subtests were biased in favor of males (Vocabulary, Arithmetic, Information, and Picture Completion). Once these biased subtests were effectively removed from the structural model, sex differences were limited to the primary factors Working Memory and Perceptual Organization. The secondary factor *g* could not account for these differences, and could therefore not be considered the source of these differences. Lynn, Fergusson, and Horwood (2005) performed similar analyses on WISC-R data obtained in a sample of 8–9-year-old children from New Zealand. Boys performed significantly better than girls on Objects Assembly, Block Design, Information, and Vocabulary, while girls performed better than boys on the subtest Coding. Again, these differences on the level of the subtests could not be attributed to differences in *g*.

The exact relation between these two methods of studying group differences, and the advantages of MG-

CMSA over the method of correlated vectors, have been discussed in detail elsewhere (e.g., Ashton & Lee, 2005; Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke, Dolan, & Kelderman, 2001; Lubke et al., 2003; Millsap, 1997). Advantages of MG-CMSA over the method of correlated vectors include higher sensitivity to model violations, and the possibility to test more specific and competing hypotheses. We therefore chose to use MG-CMSA in the following analyses. Below we first outline the MG-CMSA models that we use to investigate the source(s) of sex difference on the Dutch WAIS-III. We fit both first-, and second-order factor models, with the second-order factor representing *g*. We then present the results of the MG-CMSA analyses. Finally we discuss the results in the light of former studies.

2. Method

2.1. Subjects

Subjects were recruited from the Netherlands Twin Registry (Boomsma, 1998). All subjects participated in a large and ongoing project on the genetics of cognition. Data were available of 262 families, the total group consisting of 522 subjects (228 males and 294 females) aged 18 to 46 years. The sizes of the families ranged between sibships of size 1 ($N=13$), size 2 ($N=149$), size 3 ($N=88$), and size 4 ($N=12$). Data were available from 231 complete twin pairs (37 monozygotic males, 46 monozygotic females, 24 dizygotic males, 41 dizygotic females, and 41 dizygotic opposite twins), and, due to missingness, 10 single twins. In addition, data were available from 128 siblings, and 2 triplets.

Table 1 shows the percentages of attained educational level for males and females separately. Four educational categories were distinguished (following e.g., Schrijvers, Stronks, van de Mheen, & Mackenbach, 1999; Stronks, van de Meehn, & Mackenbach, 1997): primary education only (1), lower general and vocational education (2), intermediate vocational education, and intermediate/higher general education (3), and higher vocational education, college and university (4). Mean educational level proved equal for males and females ($F<1$, ns).

Table 1
Attained educational level for males and females in percentages

	Males	Females
Level 1	.4%	1.4%
Level 2	12.5%	14.9%
Level 3	54.5%	46.9%
Level 4	32.6%	36.8%
Mean	3.19	3.19
(SD)	(.66)	(.73)

Level 1=primary school only; Level 2=lower general, and vocational education; Level 3=intermediate vocational, and intermediate/higher general education; Level 4=higher vocational, college and university.

2.2. Tests

Psychometric IQ was measured with an abridged version of the Dutch adaptation of the WAIS-IIIR (Wechsler, 1998), the Dutch WAIS-III (WAIS-III, 1997). The following ten (of the original fourteen) subtests were administered. The subtest *Information* (IF) measures general knowledge and information gathered from daily life. In the subtest *Similarities* (SIM), subjects are asked to indicate the similarity between two verbally presented concepts. In the *Vocabulary* (VOC) subtest, subjects are asked to verbally describe the meaning of a given term. The subtest *Arithmetic* (AR) requires subjects to solve as many verbally presented arithmetic problems as possible within a given time limit without the use of paper or pencil. *Letter–Number Sequencing* (LN) requires subjects to repeat a random sequence of up to eight letters and numbers, and to put them into numerical and alphabetical order. In the subtest *Block Design* (BP), subjects are required to copy red and white patterns using red and white blocks within a given time limit. In *Matrix Reasoning* (MX), subjects are asked to select the missing part of a logical sequence out of five alternatives. In *Picture Completion* (PC), subjects have to indicate what essential part has been omitted from a picture. The subtest *Digit–Symbol Substitution* (SUB) requires subjects to replace numbers with specified symbols as fast and accurately as possible. *Copying* (CO) requires subjects to copy as many symbols as possible within a given time limit.[1]

Following the WAIS-III guidelines (WAIS-III, 1997), four dimensions are distinguished: Verbal Comprehension (VC, indicated by *Information*, *Similarities* and *Vocabulary*), Working Memory (WM, indicated by *Arithmetic* and *Letter–Number Sequencing*), Perceptual Organisation (PORG, indicated by *Block Design*, *Matrix Reasoning* and *Picture Completion*), and Perceptual Speed (PSPD, indicated by *Digit–Symbol Substitution* and *Copying*). The presence of these four dimensions was recently confirmed by a re-analysis of the WAIS-III manual data by Deary (2001).

[1] This test was not part of the final WAIS-III battery.

2.3. Statistical analyses

Multi-group confirmatory covariance and means structure analysis (MG-CMSA) was used to study sex differences in means and covariances within the common factor model. Before we examine the exact nature of sex differences in the common factors, we need to establish that the Dutch WAIS-III is measurement invariant with respect to sex. Measurement invariance with regard to sex means that the distribution of the observed scores on an indicator (i.e., subtest) of case i (y_i) depends only on his or her value on the latent variable (η_i) and not on sex: $f[y_i|\eta_i, \text{sex}]=f[y_i|\eta_i]$ (Mellenbergh, 1989; Meredith, 1993). Given normally distributed data, measurement invariance can be defined in terms of the means and the variances of the y_i given η_i, For the means, for instance, this would imply $E[y_i|\eta_i, \text{sex}]=E[y_i|\eta_i]$. In terms of the common factor model, the requirement for measurement invariance translates into specific constraints on the model parameters over groups, i.e., in this case, over sex (Meredith, 1993). Viewing factor analysis as essentially involving the linear regression of an observed variable (indicator) on latent variable (common factor), the requirements are 1) equality of the factor loadings, or regression coefficients, 2) equality of the observed means, or intercepts (while factor means are allowed to differ over groups), and 3) equality of the residual variances, i.e., the variance not attributable to the factor. If these equalities hold to reasonable approximation, the function relating the observed variables to the latent variables can be considered identical over sex, and we can interpret individual differences and group differences in terms of differences on the common factors. In factor analytic terms, the above constraints give rise to strict factorial invariance (SFI), which in turn can be interpreted as measurement invariance (in this context of the factor model; see Meredith, 1993). Failure of strict factorial invariance can have a variety of causes, including measurement bias. We refer the reader to Lubke et al. (2003) and Dolan et al. (2004) for discussion of the concept of SFI.

Below, we discuss the sequence of models fitted first to establish the presence of measurement invariance, and second to establish the source of the observed mean sex differences on IQ subtest.

2.4. Sex difference: model-fitting strategy

Adhering to the theoretical factor structure of the WAIS-III, four first-order factors are distinguished: Verbal Comprehension (VC), Working Memory (WM), Perceptual Organization (PORG), and Perceptual Speed (PSPD). In order to test the hypothesis that males and females differ with respect to g, we first need to establish measurement invariance among the 10 WAIS-III subtests. We begin by fitting a series of first-order factor models (F_1–F_4) with correlated factors (as illustrated in Fig. 1), in order to test for measurement invariance across sex. Subsequently, we introduce the second-order factor 'g' (i.e., hierarchical factor model, as illustrated in Fig. 3), and fit a series of second-order factor models (S_1–S_4) to test the hypothesis that g is the source of observed mean differences between males and females.

We start by fitting the least constrained first-order factor model, i.e., the model that tests for *configural invariance* (Horn & McArdle, 1992; Widaman & Reise, 1997). We denote this model F_1. In this model, the configuration of the regressions of the 10 subtests on the common factors is identical in males and females. The configuration is based on the expected Dutch WAIS-III factor structure (see Fig. 1). However, the exact values of the factor loadings are allowed to differ across sex. The means are included in this model as well, yet the means are unconstrained; i.e., in both males and females, the observed means (i.e., the means of the 10 subtests) are estimated freely. Model F_1 serves as a baseline model by which subsequent more constrained models are tested. We do not further consider the specific fit of this baseline model, as the assumed factor structure of the Dutch WAIS-III has been confirmed elsewhere (e.g., Deary, 2001). Note that we fixed the four factor variances to 1 in both groups, in order to establish identification of the model (e.g., see Bollen, 1989).

In model F_2 we test for *metric invariance* (Horn & McArdle, 1992; Widaman & Reise, 1997). Here we introduce the constraint that the first-order factor loadings are identical in males and females. This constraint renders fixation of the first-order factor variances in both groups superfluous, so in model F_2, these parameters are estimated freely in the females. Identical factor loadings over sex are a necessary condition to compare males and females with respect to their common factors. If the fit of model F_2 is not significantly worse than the fit of model F_1, metric invariance is considered to be tenable.

In model F_3, the model for *strong factorial invariance* (Horn & McArdle, 1992; Widaman & Reise, 1997), we introduce the structure for the means. Here, the observed means of the 10 subtests are constrained to be identical across sex, and the means of the four factors are estimated. If the fit of this model is not

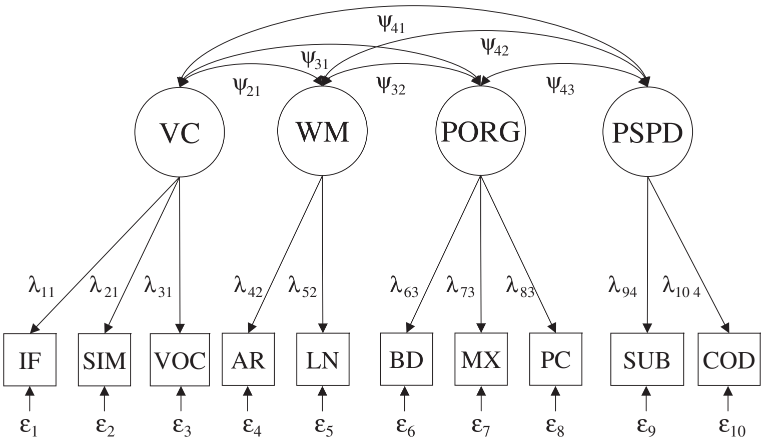

Fig. 1. First-order factor model as based on the theoretical factor structure of the WAIS-III, where the λs denote the regressions of the 10 subtests on the factors, the Ψs denote the correlations between the factors, and the εs denote those parts of the variances in the subtests that are *not* predicted by the factors, i.e., the residual variances.

significantly worse than the fit of model F_2, the assumption that expected values of the subtest scores depend only on the latent factor scores, and not on sex (i.e., $E[y_i|\eta_i, \text{sex}]=E[y_i|\eta_i]$), is tenable. It is however not possible to estimate the factor means in males and females simultaneously. Following Sörbom (1974), we fix the factor means of one group, in this case the males, to zero, and estimate the factor means of the females. Modeled as such, the male function as a reference group, and the estimated factor means of the females should be interpreted as deviations from the means of the males. This model tests whether differences in observed subtest means between males and females, can actually be attributed to differences between males and females in means on the four primary factors VC, WM, PORG and PSPD. If model F_3 is tenable (i.e., if the fit of model F_3 is not significantly worse than the fit of model F_2), the four latent factors account for the differences in observed means between males and females. In that case, we can meaningfully compare males and females with respect to their first-order factor means, i.e., with respect to the means of the broad primary factors of intelligence. However, if model F_3 does not hold, some, or all, differences in observed subtest means between males and females cannot be accounted for by the first-order factors, and can thus not be attributed to differences in the broad primary factors of intelligence. In that case we conclude that (some) subtest may be biased with respect to sex. This model for the means is illustrated in Fig. 2.

Finally, in model F_4 we test for *strict factorial invariance* (Horn & McArdle, 1992; Meredith, 1993;

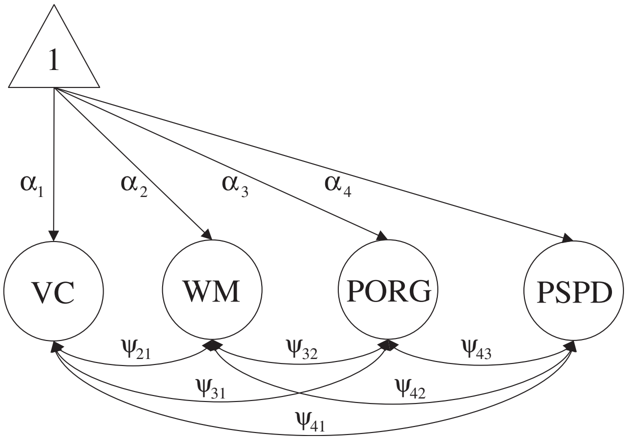

Fig. 2. Model for the means as introduced in model F_3. Note that the triangle represents a unit constant (notation introduced by John J. McArdle, see e.g., McArdle and Epstein, 1987). The α-parameters are fixed to zero in the male population, and are estimated in the female population, as such denoting the deviations from the means of the males. The Ψs denote the correlations between the factors.

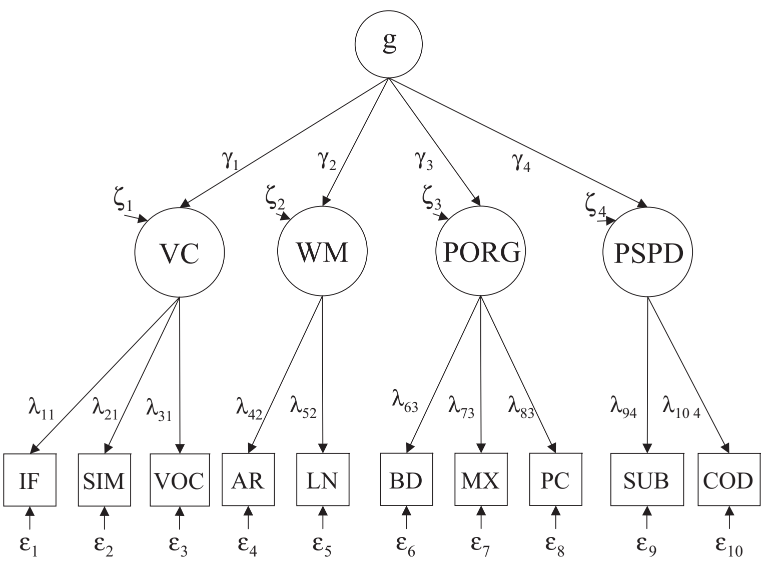

Fig. 3. The hierarchical factor model as based on the theoretical factor structure of the WAIS-III, where the λs denote the regressions of the 10 subtests on the factors, the γs denote the regressions of the first-order factor on the second-order factor g, and the ζs and εs denote those parts of the variances in the subtests and first-order factors that are *not* predicted by the first-order factors and the second-order factor, respectively.

Widaman & Reise, 1997). In this model the residual variances of the 10 subtests are constrained to be equal across sex. If model F_4 is tenable (i.e., does not fit significantly worse than model F_3), we conclude that all differences between males and females concerning the means and variances of the 10 subtests can be accounted for by the four first-order factors.

Only if factorial invariance can be established, i.e., if either model F_3 or model F_4 is tenable, do we introduce the second-order factor g. Whether this second-order factor is introduced in model F_3 or F_4 depends on the tenability of the constraints introduced in F_4. The fitted model is illustrated in Fig. 3. In this first hierarchical factor model, which will be denoted S_1, we simply test whether a hierarchical factor model is tenable in both males and females. At this point, the second-order factor loadings are allowed to differ across sex. Model S_1 simply tests whether all relations between the first-order factors VC, WM, PORG and PSPD, can be accounted for by 1 second-order factor g. Note that the means of the second-order factor are fixed to zero in both groups, while the first-order factor means are estimated for females, and fixed to zero for males (as in F_3 and F_4). This model for the means is illustrated in Fig. 4. Note also that we fixed the second-order factor variances to 1 in both groups, in order to establish identification of the model.

In model S_2, we test whether the values of the second-order factor loadings are equal across sex. Model S_2 implies that the loadings of the first-order factors on

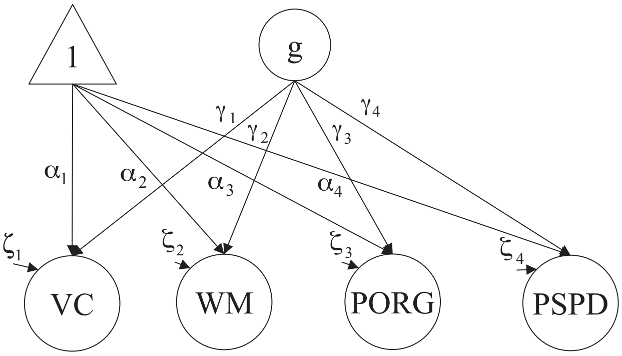

Fig. 4. Model for the means as introduced in model S_3. Note that the triangle represents a unit constant. The α-parameters are fixed to zero in the male population, and are estimated in the female population, as such denoting the deviations from the means of the males. The γs denote the regressions of the first-order factor on the second-order factor g, and the ζs denote the residual variances of the first-order factors.

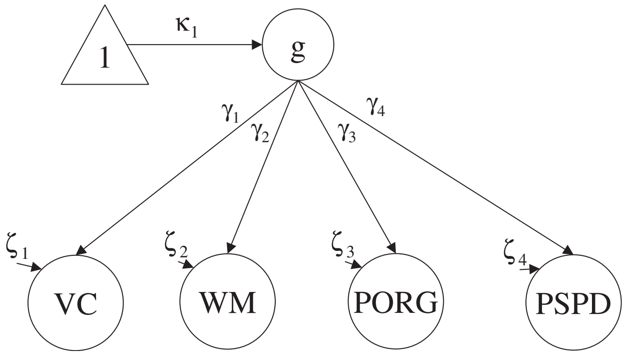

Fig. 5. Model for the means as introduced in model S_5. Note that the triangle represents a unit constant. The κ-parameter is fixed to zero in the male population, and is estimated in the female population, as such denoting the deviation from the mean of the males. The γs denote the regressions of the first-order factor on the second-order factor g, and the ζs denote the residual variances of the first-order factors.

the second-order factor are equal in males and females. This constraint renders fixation of the second-order factor variances in both groups superfluous, so in model S_2, this parameter is estimated freely in the females.

In model S_3, we constrain the first-order factor means to be equal across sex. In practice, this implies fixing the first-order factor means of the females to zero, while freely estimating the second-order factor mean. As in model F_3, it is impossible to simultaneously estimate the second-order means of the males and the females. We therefore constrain the second-order factor mean of the males to zero, and estimate the mean of the females. So again, the males function as a reference group, and the second-order factor mean of the females should be interpreted as deviation of the mean of the males. This model for the means is illustrated in Fig. 5. Note that in this model, mean differences between males and females are described entirely in terms of the means of the second-order factor g. If model S_3 holds (i.e., does not fit significantly worse than model S_2), we conclude that mean difference between males and females on the level of the subtests and the first-order factors, can completely be accounted for by differences in g. However, if the model S_3 does not fit, males and females may differ with respect to the broad primary factors of intelligence (i.e., specific abilities) but these differences cannot be attributed to differences in general intelligence. Note that it is possible in principle to estimate $p-1$ first-order factor means (where p is the total number of first-order factors in the model) *in addition to* the second-order factor mean. In that case, one allows that some first-order factor mean differences between the groups are accounted for by the second-order factor, while some are not.

Finally, in model S_4 we test whether the mean of the second-order factor g is equal across sex. If model S_4 does not fit significantly worse than S_3, we conclude that males and females do not differ with respect to g. If the fit of model S_4 is significantly worse than the fit of model S_3, it is concluded that males and females differ with respect to general intelligence or g. In Appendix A, models F_1 to F_4 and S_1 to S_4 are presented in matrix notation.

2.5. Estimation

Since the sample consisted of sibships of varying size, and scores on some subtests were missing,[2] models were fitted to the raw data (rather than directly to the covariance matrices) using Full Information Maximum Likelihood (FIML). FIML uses all observed values (e.g., Arbuckle, 1996; Finkbeiner, 1979). The Mx program (Neale, Boker, Xie, & Maes, 2003) was used for all analyses, as this program can handle such incomplete and unbalanced data. Mx calculates twice the negative loglikelihood for each model (-2LL). The difference in the -2LL of two nested models (were the nested model is the more restricted model) is asymptotically distributed as χ^2 if the restrictions are tenable (Azzelini, 1996). The number of degrees of freedom for the test comparing two nested models (the so-called χ^2 difference test, χ^2_{diff} test from here on), is equal to the difference in the number of parameters being estimated in the two compared models. The more restricted model is accepted as the preferred model, if its fit is not significantly worse than the fit of the more lenient model, i.e., if the χ^2_{diff} test is not significant. When comparing the fit of two nested

[2] Missingness was present on the following tests. COD: 4 males, 3 females; SUB: 3 males, 2 females; PC: 2 males, 2 females; BD: 2 males, 2 females; MX: 1 male. Taken together there were 21 missing values, which is less than .5% of the data.

models, a criterion level of $\alpha = .01$ was considered reasonable given the complexity of the models.

Because the focus of this paper is on sex differences, we are not concerned with the correlations among the family members. Yet, as siblings share both genetic and environmental effects, and individual differences in cognitive ability tests are known to be quite heritable (e.g., Daniels, Devlin, & Roeder, 1997), correlations between family members are expected to be present. In some former studies, data gathered within families were treated as independent observations (e.g., Jensen, 1994; Pennington et al., 2000; Wickett, Vernon, & Lee, 2000). Treating the cases as independently distributed is however not advisable in the present context, because it results in incorrect χ^2 values. We therefore fitted the factor models to the individual sibships (ranging from 1 to 4 sibs), rather than to individual cases. In so doing, we were able to estimate the covariances between the monozygotic twins, the dizygotic twins, and between the ordinary sibs, respectively.[3] These sets of covariances are considered as nuisance parameters, which are not of interest here. The actual factor model was thus fitted to the male and female data while taking into account the inherent dependency of the data. The procedure is outlined in Appendix B, and the details of the Mx specification are provided in Van der Sluis and Dolan (http://users.fmg.uva.nl/cdolan/).

3. Results

All following analyses were performed on the raw, unstandardized IQ scores. Correlations between raw scores and standardized norm scores ranged between .95 and .99.

3.1. Preliminary analyses

Table 2 shows the means and standard deviations of the norm scores for males and females separately, as well as the effect sizes (Cohen's *d*) for each subtest, calculated as the differences between the means of the males and females ($\mu_m - \mu_f$) divided by their pooled standard deviation. The table also shows the composite scores Verbal IQ (VIQ, based on INF, SIM, VOC and AR), Performance IQ (PIQ, based on BD, MX, PC, and SUB), and Full scale IQ (FIQ, based on VIQ and PIQ). In the norm population, the means of VIQ, PIQ, and FIQ are 100, with a standard deviation of 15. Table 2 shows that the means of the males and females in our sample were close to the population means. Only the mean VIQ for females was lower ($M = 93.04$, SD = 12.12).

Table 2
Means and standard deviations for males ($N = 228$) and females ($N = 294$) on the 10 WAIS-III subtests (standardized scores)

Factors	Subtests	Males			Females			d
		M	SD	*N*	*M*	SD	*N*	
VC	INF	10.74	2.92	224	8.82	2.91	291	.66
	SIM	7.35	2.28	225	6.96	1.98	292	.18
	VOC	9.97	2.80	226	9.79	2.44	292	.07
WM	AR	10.84	3.09	226	9.56	3.02	292	.42
	LN	11.61	3.18	227	11.03	2.99	294	.10
PORG	BD	9.58	2.69	228	8.88	2.75	294	.26
	MX	10.84	1.71	228	10.46	2.04	294	.20
	PC	11.01	2.56	228	10.45	2.38	294	.23
PSPD	SUB	9.74	2.67	228	11.73	2.90	294	−.71
	CO	26.60	3.34	228	18.55	3.15	294	−.19
VIQ		98.62	13.07	228	93.04	12.12	293	.44
PIQ		102.02	10.85	228	102.86	11.89	293	−.07
FIQ		99.19	10.92	227	96.63	10.75	292	.24

VC = Verbal Comprehension, WM = Working Memory, PORG = Perceptual Organization, PSPD = Perceptual Speed, INF = Information, SIM = Similarities, VOC = Vocabulary, AR = Arithmetic, LN = Letter–Number Sequencing, BD = Block Design, MX = Matrix Reasoning, PC = Picture Completion, CO = Copying, SUB = Digit–Symbol Substitution. Following the WAIS-III manual, VIQ is based on the subtests INF, SIM, VOC and AR, PIQ is based on the subtests BP, MX, PC and SUB, and FIQ is based on the subtests INF, SIM, VOC, AR, BP, MX, PC and SUB. *d* is Cohen's effect size *d*, calculated as $(M_{\text{males}} - M_{\text{females}}) / \sigma^2_{\text{pooled}}$. Negative *d*s denote higher means for females.

To gain some insight into the mean sex differences on the level of the subtests and on the level of the composite scores, multivariate analyses of variance (MANOVA), and *t*-tests for independent samples were conducted in SPSS. Because of the mutual dependence of the data, these analyses were not performed on all 522 cases simultaneously. Rather, the analyses were conducted separately for all first (83 males, 111 females), and all second (81 males, 108 females) members of each family for who complete data were available (see Table 3 for an overview).[4] The MANOVAs for the subtests showed main effects for sex (first family members: $F(10,183) = 6.79$, $p < .000$; second family members: $F(10,178) = 7.31$, $p < .000$). Adopting an α of .01, we found that in both samples females outperformed males on SUB, and males outperformed females on INF and

[3] Although dizygotic twins and ordinary sibs are characterized by the same degree of genetic relatedness, dizygotic twin often covary more strongly, because they are matched at a variety of variables including sex and age. We therefore chose to estimate the covariance separately for dizygotic twins. Opposite sex twins were however treated as ordinary sibs.

[4] Analyses were not performed for the third and fourth members of the families, as these groups were rather small in comparison to the groups of first and second family members.

Table 3
Results of t-tests for first family members (83 males vs 111 females) and second family members (81 males vs 108 females), and paired t-tests for 38 pairs of opposite sex twins

Factors	Subtests	1st family members			2nd family members			DOS pairs		
		t	*df*	*p*	*t*	*df*	*p*	*t*	*df*	*p*
VC	INF	3.42	192	.001*	5.61	187	.000*	6.10	37	.000*
	SIM	.96	192	.338	1.36	187	.177	.64	37	.528
	VOC	− .06	192	.953	.54	187	.593	.64	37	.528
WM	AR	2.78	192	.006*	3.08	187	.002*	3.35	37	.002*
	LN	.03	192	.980	1.99	187	.048	.66	37	.511
PORG	BD	2.55	192	.012	2.56	187	.011	2.08	37	.045
	MX	3.42	192	.001*	2.01	187	.046	1.83	37	.076
	PC	1.21	192	.227	2.43	187	.016	1.58	37	.122
PSPD	SUB	− 4.88	192	.000*	− 3.03	187	.003*	− 3.20	37	.003*
	CO	− 1.36	192	.174	− 1.31	187	.194	− 1.03	36	.310
VIQ		2.57	196	.011	3.54	192	.001*	4.12	37	.000*
PIQ		− .67	196	.503	.23	192	.817	− .29	37	.77
FIQ		1.096	196	.274	2.24	192	.027	1.60	37	.12

AR. In addition, males outperformed females on MX in the sample of first family members, while this difference was not significant in the sample of second family members (p = .065). The MANOVAs for the composite scores also showed a main effect for sex (first family members: $F(3,194) = 4.37$, $p < .01$; second family members: $F(3,190) = 5.58$, $p < .01$). Adopting an α of .01, we found that males outperformed females on VIQ ($p < .01$), but differences between males and females on PIQ and FIQ were not significant (although with a p-value of .03, the difference for FIQ approached significance for the second family members).

Although the males and females in our sample did not differ with respect to their educational level, it is possible that the sex differences for SUB, INF, AR and VIQ, were the result of sex differences with regard to other SES related variables. This possibility cannot be ruled out entirely. However, our sample contained 38 complete pairs of DOS twins, i.e., 76 males and females who were perfectly matched on SES. Table 3 shows an overview of the paired t-tests, which were conducted for these 38 pairs. The results of these paired t-tests copied the results for the first and second family members: females outperformed males on SUB, but males outperformed females on INF, AR, and VIQ. We therefore consider it unlikely that the sex differences in our sample were the result of sex differences in SES.

3.2. First-order factor models

The − 2LL's of models F_1 to F_4, and the models S_1 to S_5, are presented in Table 4. All following model comparisons are summarized in Table 5. Model F1, in which only configural invariance over males and females was presumed, served as the baseline model. In model F_2, the factor loadings were constrained to be equal across sex (metric invariance). The χ^2_{diff} test, in which the fit of model F_2 was compared to the fit of model F_1, was not significant ($\chi^2_{\text{diff}}(6) = 1.31$, ns), and the factor loadings could thus be considered equal across sex. In model F_3, all intercepts were constrained to be equal across sex (strong factorial invariance), and latent differences in factor means were estimated. The fit of model F_3 was significantly worse than the fit of model F_2 ($\chi^2_{\text{diff}}(6) = 75.56$, $p < .000$), suggesting that not all sex differences on the level of the subtests could be accounted for by the first-order factors. In model F_{3a}, all intercepts were constrained to be equal across sex, except the intercept for INF. The fit of model F_{3a} was

Table 4
Results of all multi-group covariance and mean structure analyses

Model		*df*	− 2LL
F_1	Configural invariance	4962	33,259.99
F_2	Metric invariance	4968	33,261.30
F_3	Strong factorial invariance	4974	33,336.86
F_{3a}	Strong factorial invariance, bar INF	4973	33,275.41
F_{4a}	Strict factorial invariance	4983	33,291.27
F_{4b1}	$\mu_{\text{males,VC}} = \mu_{\text{females,VC}}$	4984	33,293.80
F_{4b2}	$\mu_{\text{males,WM}} = \mu_{\text{females,WM}}$	4984	33,310.48
F_{4b3}	$\mu_{\text{males,PORG}} = \mu_{\text{females,PORG}}$	4984	33,304.93
F_{4b4}	$\mu_{\text{males,PSPD}} = \mu_{\text{females,PSPD}}$	4984	33,327.78
S_1	Introduction *g*	4987	33,295.35
S_{1a}	WM equals *g* in females	4988	33,295.77
S_2	Metric invariance on 2nd order level	4990	33,295.56
S_3	Strong factorial invariance on 2nd order level	4993	33,361.06
$S_{3,\ b1}$	Strong factorial invariance on 2nd order level, bar PSPD	4992	33,297.96
S_4	$\mu_{\text{males},\ g} = \mu_{\text{females},\ g}$	4993	33,317.46

Table 5
Model comparisons

Models	χ^2	*df*	*p*
F_2 vs F_1	1.31	6	.97
F_3 vs F_2	75.56	6	<.001
F_{3a} vs F_2	14.11	5	.02
F_{4a} vs F_{3a}	15.86	10	.10
F_{4b1} vs F_{4a}	2.53	1	.11
F_{4b2} vs F_{4a}	19.21	1	<.001
F_{4b3} vs F_{4a}	13.66	1	<.001
F_{4b4} vs F_{4a}	36.51	1	<.001
S_1 vs F_{4a}	4.07	4	.40
S_{1a} vs S_1	.42	1	.52
S_2 vs S_1	.21	3	.98
S_3 vs S_2	65.50	3	<.001
S_{3a} vs S_2	2.40	2	.30
S_4 vs S_{3a}	19.50	1	<.001

not significantly worse than the fit of model F_2 (χ^2_{diff} (5)=14.11, ns). The hypothesis of strong factorial invariance thus proved tenable for all subtests, bar INF. In order to test for strict factorial invariance, the residual variances were constrained to be equal across sex in model F_{4a}. The fit of model F_{4a} was not significantly worse than the fit of model F3a (χ^2_{diff}(10)=15.86, ns).

Summarizing, we conclude that strict factorial invariance was tenable, except for the subtest INF. Thus, the sex related mean difference in the subtest INF could not be accounted for by the sex related mean difference in the factor VC. As it turns out, the female mean on the subtest INF is too low given the mean difference between males and females on the factor VC. For all other 9 subtests, the mean differences between males and females could be accounted for by the first-order factors VC, WM, PORG and PSPD.

Having established measurement invariance on the level of nine out of ten subtests, we tested whether sex differences existed with respect to the means of the four primary factors VC, WM, PORG and PSPD. Remember that these primary factor means were fixed to zero in males, while the factor means of the females were estimated as deviation of the mean of the males. The factorial means of the females are thus interpreted as deviations from the means of the males, with positive means favoring females, and negative means favoring males. Whether the factorial means of the females differ significantly from the factorial means of the males, can be established by fixing the means of the females to zero one at a time (model F_{4b1} to F_{4b4}), and comparing the fit of the resulting models, to the fit of the model in which all female means were estimated freely.

Females scored higher than males on the PSPD factor (the mean equals .67, SD=.98), but lower than males on the factors VC, WM, and PORG (the means being −.16, −.48, and −.42, with SDs of .99, .96 and 1.08, respectively). Fixing the mean of the VC factor to zero in model F_{4b1} did not result in a significant decrease in fit (χ^2_{diff} (1)=2.53, ns). For the other factors, however, fixing the means of the females to zero (F_{4b2} to model F_{4b4}) did result in significantly poorer fit (for WM: χ^2_{diff} (1)=19.21, $p<.0001$; for PORG: χ^2_{diff} (1)=13.66, $p<.001$; for PSPD: χ^2_{diff} (1)=36.51, $p<.0001$). Bearing in mind that the subtest INF was effectively removed from the structural model for the means, and does therefore not contribute to the common factor mean differences, we conclude that males scored significantly higher than females on WM and PORG, and that females scored significantly higher than males on PSPD. Sex differences on the VC factor (but excluding INF) were however absent. Table 6 contains the correlations between the four primary factors for males and females separately, and the mean differences and standard deviations between males and females on the four factors.

3.3. Second-order factor models

In model S_1, the factor *g* was introduced as a second-order factor, taking model F_{4a} as point of departure. Note that, as in model F_{4a}, the means are still modeled on the level of the first-order factors in model S1. The fit of this model was not significantly worse than that of model F_{4a} (χ^2_{diff}(4)=4.07, ns), suggesting that the second-order factor could account for the covariances between the four primary factors. In the male sample, the second-order factor explained 41% of the variance in the factor VC, 23% in PSPD, 69% in PORG, and 93% in WM. In the female sample, the second-order factor explained 42% of the variance in the first-order

Table 6
Correlations between the first-order factors Verbal Comprehension (VC), Working Memory (WM), Perceptual Organization (PORG) and Perceptual Speed (PSPD) of the WAIS-III for females (above diagonal) and males (below diagonal), and male and female means and standard deviations on the first-order factors

		VC	WM	PORG	PSPD
VC		–	.685	.556	.224
WM		.608	–	.826	.496
PORG		.539	.813	–	.383
PSPD		.273	.448	.370	–
Males	Mean	0	0	0	0
	SD	1	1	1	1
Females	Mean	−.16	−.48****	−.42***	.67****
	SD	.99	.96	1.08	.98

The means of the females should be interpreted as *deviations* from the means of the males. *** and **** denote statistical significance at α=.001 and α=.0001, respectively.

factor VC, 22% in PSPD, 65% in PORG, and 100% in WM. The estimate of the residual variance of the factor WM was negative in the females. Fixing this residual variance to zero did not result in significant loss of fit, which suggests perfect prediction of WM by the *g*-factor (Model S_{1a}, $\chi^2_{\text{diff}}(1)=.42$, ns).

In model S_2, the second-order factor loadings were constrained to be equal across sex. The fit of model S_2 was not significantly worse than the fit of model S_1 (χ^2_{diff} (3)=.21, ns), suggesting that the second-order factor loadings could be considered identical for males and females. In model S_3, the first-order factor means differences were fixed to zero, and the mean difference with respect to *g* was estimated. In model S_3, *g* is the only source of mean differences between males and females. This model was not tenable, as its fit was significantly worse than the fit of model S_2 (χ^2_{diff} (3) =65.50, $p<.000$). This suggests that the mean differences between males and females on the first-order factors could not be described as mean differences between males and females in *g*. Above we established that females outperformed males on PSPD, and that males outperformed females on WM and PORG, while sex differences were absent with respect to VC. Also, Table 6 shows that all four factors were positively correlated. Considering these varied mean differences, combined with the positive factor correlations, it is not surprising that *g* could not account for the first-order mean differences between males and females. After all, one factor cannot account for positive (PSPD) and negative (WM, PORG) mean differences simultaneously, when these occur on factors that are positively correlated. The positive sign of the mean of PSPD was thus the main cause of the misfit in S_3. In model S_{3a}, the mean of the females on PSPD was estimated freely, while the means of VC, WM and PORG were constrained to be equal across sex. This implies that the mean of the factor PSPD was effectively removed from the model of the means. The fit of this model was not significantly worse than the fit of model S_2 ($\chi^2_{\text{diff}}(2)=2.40$, ns), suggesting that the second-order factor could account for the differences between males and females on the factors VC, WM, and PORG. When the mean of the resulting/consequent second-order factor was subsequently constrained to be equal across sex in model S_5, the fit worsened significantly ($\chi^2_{\text{diff}}(1)=19.50$, $p<.000$). The mean of this second-order factor could therefore not be considered equal across sex. However, the mean of this second-order factor can also no longer be interpreted as representing *g*, as the mean of PSPD was effectively removed from the model. That is, the second-order factor had degenerated into a factor describing the advantages of males over females on WM and PORG, i.e., on a subset of the original primary factors. We therefore conclude that males and females differ with respect to the broad primary factors WM, PORG and PSPD, and that these differences were not attributable to, or could not accurately be described by differences in general intelligence, or *g*.

4. Discussion

In the present study we used multi-group covariance and mean structure analysis (MG-CMSA) to investigate whether sex differences observed on the subtests of the Dutch WAIS-III were attributable to sex difference in general intelligence (*g*). Males were found to outperform females on 3 of the 10 subtests (Information, Arithmetic, and Matrix Reasoning), while females outperformed males on 1 subtest, namely, Digit–Symbol Substitution. We first established whether the subtests of the Dutch WAIS-III were measurement invariant across sex, i.e., whether bias was absent on the level of the subtests. These analyses showed that the subtest Ifnformation was biased in favor of males. This subtest was therefore excluded from subsequent analyses in which males and females were compared with respect to their means on the primary factors Verbal Comprehension, Working Memory, Perceptual Organization, Perceptual Speed, and with respect to their means on the secondary factor for general intelligence, *g*. Males and females showed no mean differences with respect to the primary factor Verbal Comprehension. However, males had significantly higher means than females on the factors Working Memory and Perceptual Organization, while females showed higher means than males on the factor Perceptual Speed. Because some primary factors showed a male advantages, while other showed a female advantages, and all four primary factors were positively correlated in both groups, the secondary factor *g* could not account for all mean differences between males and females simultaneously. We therefore conclude that *g* is *not* the source of sex differences observed on the Dutch WAIS-III subtests and on the primary factors.

The presence of sex differences on the level of the subtests of intelligence batteries such as the WISC, the WAIS and the DAT, is not unusual, and has been reported on in former studies (e.g., Colom & Lynn, 2004; Dolan et al., submitted for publication; Lynn, 1998; Lynn, Irwing, & Cammock, 2001; Lynn et al., 2005). Especially sex differences with respect to tests measuring general knowledge, such as the subtest Information of the WAIS, are well documented (e.g., Dolan et al., submitted for publication; Jensen & Reynolds, 1983; Lynn, 1998; Lynn et al., 2001; Lynn et al., 2005).

Our finding that males outperform females on the primary factors for Working Memory and Perceptual Organization, and that males and females do not differ with respect to Verbal Comprehension, is in concordance with the results reported by Dolan et al. (submitted for publication). The females in their sample did however not show any advantages over males with respect to Perceptual Speed.

In the present data set, the secondary factor *g* proved highly correlated to the Working Memory factor, predicting 93% of the variance in males, and 100% of the variance in females. The finding that *g* is strongly, or even perfectly, related to one of the first-order factors, is not unique. *g* has been found to relate strongly to broad primary factors such as verbal intelligence (G_v), fluid intelligence (G_f), crystalized intelligence (G_c), quantitative ability (G_q), and retrieval ability (G_r) (e.g., Bickley, Keith, & Wolfle, 1995; Dolan, 2000; Gustafsson, 1984; Undheim & Gustafsson, 1987), as well as to factors for working memory (e.g., Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004; Conway, Cowan, Bunting, Therriault, & Monkoff, 2002; Kyllonen & Christal, 1990). With respect to the WAIS-III and the WAIS-IIIR, the relation between *g* and the four primary factors appears to be population dependent. For example, in their re-analysis of the data gathered from the Spanish standardization sample, Dolan et al. (submitted for publication) report correlations between WM and *g* of .66 and .62 in males and females, respectively. In contrast, in a re-analysis of the American standardization sample (i.e., the WAIS-III manual data) Deary (2001) reports a correlation of .95 between WM and *g* (calculated across sex). The fact that the correlations between *g* and the broad primary factors vary across test batteries and across samples (i.e., across countries) seems to indicate that the extracted secondary factor *g* is not always the same. The interpretation, strength, value and importance of *g*, and its relation to the primary factors, depend on the specific choice of subtests in the battery, and on the population from which the data are gathered (see also Horn & Noll, 1997).

The present study illustrated the advantages of MG-CMSA in the investigation of mean group differences: MG-CMSA allows for explicit, integrated model specification, goodness of fit testing, and explicit comparison of competing hypotheses within the model framework. The advantages of MG-CMSA over the method of correlated factors were shortly mentioned in the introduction, and are discussed in more detail elsewhere (e.g., Ashton & Lee, 2005; Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001; Lubke et al., 2003; Millsap, 1997). In the present data, the method of correlated vectors yielded a correlation (Spearman's rho) between the *g*-loadings (obtained in a Schmid–Leiman factor analysis) and standardized mean differences of .68. Jensen (1998) reported an average value of about .60 for Black–White differences, and interpreted this as strong support for Spearman's hypothesis. The value of .68, as found for the present data, may thus suggest that *g* is the main source of sex differences. The results of the MG-CMSA clearly show that this conclusion is hardly tenable: the configuration of factor mean differences (see Table 6) *in combination with* the positive manifold among the four primary factors (i.e., implying that the second order factor loadings on *g* are all positive), renders the hypothesis that *g* is the main source of sex mean differences unlikely. This again illustrates the point that the method of correlated vectors is too blunt an instrument to reliably unveil the nature of latent mean differences (see also Ashton & Lee, 2005; Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001). We note that Lynn (1999) adopted a different approach. He calculated the sum of the primary common factor scores, and presented this summation as a measure of the mean *g*. Application of this procedure to the present data would imply summation of the differences between males and females in their first-order factor means. The mean difference in *g* between males and females would then turn out to be $+.67-.16-.48-.42=-.39$ (see Table 6), and we would conclude that males are endowed with higher general intelligence than females. However, to establish that these first order mean differences are a function of *g*, we have to (1) actually fit the model that represents this hypothesis (model S_3 above), and (2) compare the results of this models and competing models both with respect to goodness of fit and interpretability. Calculating the sum of the differences between males and females in their first-order factor means may create the impression that *g* is the source of the differences, but it cannot be put forward as evidence that this is indeed the case. Specifically, Jensen's statement that significant group differences on a collection of IQ subtests are not necessarily attributable to differences in *g* (Jensen, 1998) holds equally for significant group differences on first-order factors of intelligence.

Summarizing, the present study showed that males and females do differ with respect to specific cognitive abilities, but that *g* could not be put down as the source of these differences. Finally, we emphasize that the establishment of measurement invariance is not only important in the investigation of latent mean differences between groups, as we have discussed in detail here. Measurement invariance is also important in investigating group differences in the relationships of latent variables such as the primary cognitive factors or *g* with external variables such as brain volume, educational attainment, and age.

Appendix A

In the present Appendix, we present the models discussed above in matrix notation of the LISREL program (Jöreskog & Sörbom, 2001). Let $\boldsymbol{y}_{ij}$ denote the observed *r*-dimensional random column vector containing the observed scores of subject *j* in population *i*. The following common factor model is assumed to hold for observation $\boldsymbol{y}_{ij}$:

$$\boldsymbol{y}_{ij} = \boldsymbol{\nu}_{yi} + \boldsymbol{\Lambda}_i \boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{A1}$$

where $\boldsymbol{\eta}_{ij}$ is a *p*-dimensional vector of first-order factors, and $\boldsymbol{\varepsilon}_{ij}$ is a *r*-dimensional vector of residuals. Note that these residuals contain both measurement error and systematic error. $\boldsymbol{\Lambda}_i$ is a $r \times p$ dimensional matrix of first-order factor loadings, which can be interpreted as regression coefficients, in the regression of the subtest scores $\boldsymbol{y}_{ij}$ on the first-order factors $\boldsymbol{\eta}_{ij}$. $\boldsymbol{\nu}_{yi}$ is a *r*-dimensional vector of intercepts in this same regression.

In a hierarchical second-order factor model, where the first-order factors are themselves regressed on a second-order factor (i.e., on *g*), the following model is assumed to hold for the first-order factors $\boldsymbol{\eta}_{ij}$:

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\alpha}_i + \boldsymbol{\Gamma}_i \boldsymbol{\xi}_{ij} + \boldsymbol{\zeta}_{ij} \tag{A2}$$

where $\boldsymbol{\xi}_{ij}$ is a *q*-dimensional vector of second-order factors (as we have only 1 second-order factor, this is a scalar), $\boldsymbol{\Gamma}_i$ is the $p \times q$ dimensional matrix of second-order factor loadings, which can be interpreted as the regression coefficients in the regression of the first-order factors $\boldsymbol{\eta}_{ij}$ on the second-order factor $\boldsymbol{\xi}_{ij}$. The $(p \times 1)$ vector $\boldsymbol{\alpha}_i$ can be interpreted as a vector of regression intercept terms. The *p*-dimensional vector of $\boldsymbol{\zeta}_{ij}$ contains residuals. So, for a hierarchical factor model with first- and second-order factors, the complete common factor model for observation $\boldsymbol{y}_{ij}$ is given as:

$$\boldsymbol{y}_{ij} = \boldsymbol{\nu}_{yi} + \boldsymbol{\Lambda}_i\left(\boldsymbol{\Gamma}_i \boldsymbol{\xi}_{ij} + \boldsymbol{\zeta}_{ij}\right) + \boldsymbol{\varepsilon}_{ij}. \tag{A3}$$

We further assume that $\boldsymbol{\varepsilon}_{ij}$ is independent of $\boldsymbol{\zeta}_{ij}$ and $\boldsymbol{\xi}_{ij}$, and that $\boldsymbol{\zeta}_{ij}$ and $\boldsymbol{\xi}_{ij}$ are independent. Given these assumptions, and given that the observed variables $\boldsymbol{y}_{ij}$ are normally distributed $\boldsymbol{y}_{ij} \sim \mathrm{N}_r(\mu_i, \Sigma_i)$, the model implied covariance matrix Σ_i for group *i* is given as:

$$\boldsymbol{\Sigma}_i = \Lambda_i(\boldsymbol{\Gamma}_i \boldsymbol{\Phi}_i \boldsymbol{\Gamma}_i t + \boldsymbol{\Psi}_i)\Lambda_i^t + \boldsymbol{\Theta}_i, \tag{A4}$$

where $\boldsymbol{\Psi}$ denotes the $p \times p$ covariance matrix of the first order factors, $\boldsymbol{\Phi}$ denotes the $q \times q$ covariance matrix of the second-order factors, and $\boldsymbol{\Theta}$ is the $r \times r$ covariance matrix of the residuals. The superscript *t* denotes transposition. The model implied vector for the means is given as:

$$\boldsymbol{\mu}_i = \boldsymbol{\nu}_i + \Lambda_i \boldsymbol{\alpha}_i + \Lambda_i \boldsymbol{\Gamma}_i \boldsymbol{\kappa}_i, \tag{A5}$$

where $\boldsymbol{\nu}_i$ denoted the *r*-dimensional vector of observed intercepts, $\boldsymbol{\alpha}_i$ denotes the *p*-dimensional vector of first-order factor means, and $\boldsymbol{\kappa}_i$ the *q*-dimensional vector of second-order factor means.

We fit the following sequence of first-order factor models in the male (subscript m) and female (subscript f) samples. Model F_1 can be denoted as:

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_{\mathrm{m}} \boldsymbol{\Psi}_{\mathrm{m}} \boldsymbol{\Lambda}_{\mathrm{m}}^t + \boldsymbol{\Theta}_{\mathrm{m}}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{\mathrm{m}},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_{\mathrm{f}} \boldsymbol{\Psi}_{\mathrm{f}} \boldsymbol{\Lambda}_{\mathrm{f}}^t + \boldsymbol{\Theta}_{\mathrm{f}}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{\mathrm{f}}. \tag{A6}$$

In model F_1, $\boldsymbol{\Lambda}_{\mathrm{m}}$ and $\boldsymbol{\Lambda}_{\mathrm{f}}$ have the same configuration of first-order factor loadings, as based on the expected WAIS-III factor structure, but equality constraints over sex are absent. Note that the intercepts $\boldsymbol{\nu}_{\mathrm{m}}$ and $\boldsymbol{\nu}_{\mathrm{f}}$ are estimated freely and the factor means are fixed to zero.

In model F_2, the first-order factor loadings are constrained to be equal across sex, i.e., $\boldsymbol{\Lambda}_{\mathrm{m}}=\boldsymbol{\Lambda}_{\mathrm{f}}=\boldsymbol{\Lambda}_*$:

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_* \boldsymbol{\Psi}_{\mathrm{m}} \boldsymbol{\Lambda}_*^t + \boldsymbol{\Theta}_{\mathrm{m}}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{\mathrm{m}},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_* \boldsymbol{\Psi}_{\mathrm{f}} \boldsymbol{\Lambda}_*^t + \boldsymbol{\Theta}_{\mathrm{f}}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{\mathrm{f}}. \tag{A7}$$

where $\boldsymbol{\Lambda}_*$ denotes the matrix with constrained first-order factorloadings.

In model F_3, the structure for the means is introduced. That is, the vectors with observed means are constrained to be identical across sex, i.e., $\boldsymbol{\nu}_{\mathrm{m}}=\boldsymbol{\nu}_{\mathrm{f}}=\boldsymbol{\nu}_*$. As the factor means cannot be estimated simultaneously in males and females, the factor means of the males are fixed to zero, while the factor means of the females are estimated freely.

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_* \boldsymbol{\Psi}_{\mathrm{m}} \boldsymbol{\Lambda}_*^t + \boldsymbol{\Theta}_{\mathrm{m}}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_*, \\
&\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_* \boldsymbol{\Psi}_{\mathrm{f}} \boldsymbol{\Lambda}_*^t + \boldsymbol{\Theta}_{\mathrm{f}}, \quad \text{and} \\
&\quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_* + \boldsymbol{\Lambda}_*(\boldsymbol{\alpha}_{\mathrm{m}} - \boldsymbol{\alpha}_{\mathrm{f}}) = \boldsymbol{\nu}_* + \boldsymbol{\Lambda}_* \boldsymbol{\alpha}_{\triangle},
\end{aligned} \tag{A8}$$

where $\boldsymbol{\alpha}_{\triangle}$denotes $(\boldsymbol{\alpha}_{\mathrm{m}}-\boldsymbol{\alpha}_{\mathrm{f}})$, i.e., the difference between the factor means of the males and the females.

In model F4, the residual variances are constrained to be equal across sex, i.e., $\boldsymbol{\Theta}_{\mathrm{m}}=\boldsymbol{\Theta}_{\mathrm{f}}=\boldsymbol{\Theta}_{*}$:

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_{*}\boldsymbol{\Psi}_{\mathrm{m}}\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{*},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_{*}\boldsymbol{\Psi}_{\mathrm{f}}\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{*} + \boldsymbol{\Lambda}_{*}\boldsymbol{\alpha}_{\triangle}. \tag{A9}$$

In model S_1, the second-order factor is introduced in either model F_3 or F_4, depending on whether model F_4 is tenable. Let us suppose strict factorial invariance (i.e., model F_4) is tenable. Model F_4 then functions as a baseline model in which the second-order factor g is introduced. In model S_1, $\boldsymbol{\Gamma}_{\mathrm{m}}$ and $\boldsymbol{\Gamma}_{\mathrm{f}}$ have the same configuration of second-order factor loadings, but equality constraints over sex are absent.

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{\mathrm{m}}\boldsymbol{\Phi}_{\mathrm{m}}\boldsymbol{\Gamma}_{\mathrm{m}}^{t} + \boldsymbol{\Psi}_{\mathrm{m}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{*},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{\mathrm{f}}\boldsymbol{\Phi}_{\mathrm{f}}\boldsymbol{\Gamma}_{\mathrm{f}}^{t} + \boldsymbol{\Psi}_{\mathrm{f}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{\mathrm{f}} + \boldsymbol{\Lambda}_{*}\boldsymbol{\alpha}_{\triangle}. \tag{A10}$$

In model S_2, we test whether the second-order factor loadings can be considered equal across sex, i.e., $\boldsymbol{\Gamma}_{\mathrm{m}}=\boldsymbol{\Gamma}_{\mathrm{f}}=\boldsymbol{\Gamma}_{*}$:

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{m}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{\mathrm{m}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{*},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{f}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{\mathrm{f}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{*} + \boldsymbol{\Lambda}_{*}\boldsymbol{\alpha}_{\triangle}. \tag{A11}$$

In model S_3, the first-order factor means are constrained to be equal across sex, i.e., $\boldsymbol{\alpha}_{\mathrm{m}}=\boldsymbol{\alpha}_{\mathrm{f}}$. As the means of the males $(\boldsymbol{\alpha}_{\mathrm{m}})$ were fixed to zero, this implies fixing the means of the females $(\boldsymbol{\alpha}_{\mathrm{f}})$ to zero as well. Due to lack of identification, it is impossible to estimate the second-order mean in males $(\boldsymbol{\kappa}_{\mathrm{m}})$ and females $(\boldsymbol{\kappa}_{\mathrm{f}})$, simultaneously, so $\boldsymbol{\kappa}_{\mathrm{m}}$ is fixed to zero, while $\boldsymbol{\kappa}_{\mathrm{f}}$ is estimated:

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{m}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{\mathrm{m}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{*},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{f}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{\mathrm{f}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{*} + \boldsymbol{\Lambda}_{*}\boldsymbol{\Gamma}_{*}(\kappa_{m-}\kappa_{\mathrm{f}}), = \boldsymbol{\nu}_{*} + \boldsymbol{\Lambda}_{*}\boldsymbol{\Gamma}_{*}\boldsymbol{\kappa}_{\triangle}, \tag{A12}$$

where $\boldsymbol{\kappa}_{\triangle}$ denotes $\boldsymbol{\kappa}_{\mathrm{m}}-\boldsymbol{\kappa}_{\mathrm{f}}$, i.e., the difference between the second-order factor mean of the males and the females.

In principle it is possible to estimate $p-1$ first-order factor means $\boldsymbol{\alpha}$ (where p is the total number of first-order factors in the model), besides estimation of the mean of the second-order factor $\boldsymbol{\kappa}$. In that case, one creates a model in which some first-order factor mean differences between the groups are accounted for by the second-order factor, while some are not:

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{m}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{\mathrm{m}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{*},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{f}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{\mathrm{f}}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{*} + \boldsymbol{\Lambda}_{*}\boldsymbol{\alpha}_{\triangle} + \boldsymbol{\Lambda}_{*}\boldsymbol{\Gamma}_{*}\boldsymbol{\kappa}_{\triangle}. \tag{A13}$$

In model S_4, the mean of the second-order factor g is constrained to be equal across sex. As the mean of the males $(\boldsymbol{\kappa}_{\mathrm{m}})$ was fixed to zero, this implies fixing the mean of the females $(\boldsymbol{\kappa}_{\mathrm{f}})$ to zero as well:

$$\boldsymbol{\Sigma}_{\mathrm{m}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{m}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{*}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{m}} = \boldsymbol{\nu}_{*},$$

$$\boldsymbol{\Sigma}_{\mathrm{f}} = \boldsymbol{\Lambda}_{*}\left(\boldsymbol{\Gamma}_{*}\boldsymbol{\Phi}_{\mathrm{f}}\boldsymbol{\Gamma}_{*}^{t} + \boldsymbol{\Psi}_{*}\right)\boldsymbol{\Lambda}_{*}^{t} + \boldsymbol{\Theta}_{*}, \quad \text{and} \quad \boldsymbol{\mu}_{\mathrm{f}} = \boldsymbol{\nu}_{*}. \tag{A14}$$

Appendix B

To explain how we dealt with dependence in the data, we consider a simple situation in which we have observed data in 4 individuals: $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, $\boldsymbol{y}_3$, and $\boldsymbol{y}_4$. Each data vector $\boldsymbol{y}_i$ comprises p observed variables. Suppose that we fit some covariance and mean model to the data using raw data maximum likelihood estimation. Let S^* and m^* denote the model implied covariance and mean structure, respectively. E.g., S^* may be a single common factor model. If the 4 cases are independent, their individual contributions to the loglikelihood function may be simply added:

$$LogL = \sum_{i=1}^{4} \log\Big(|S^*|^{-1/2}(2\pi)^{-p/2} \times \exp\big[\big(-1/2(y_i - m^*)^t S^{*-1}(y_i - m^*)\big)\big]\Big) \tag{B1}$$

Note that in the case of multi-group analyses, in which males and females are modeled separately, the above function is fitted separately for males and females, and separate covariance and mean structures are constructed, i.e., S_{m}^* and m_{m}^*, and S_{f}^* and m_{f}^*, respectively.

In the present data, dependence of the 4 cases arises because they are members of the same family. To accommodate the dependence, we treat each family as a case, rather than treating each individual as a case. We achieve this by specifying the following covariance matrix and mean vector for each family separately:

$$S^{\circ} = \begin{bmatrix} S* & & & \\ S_{21} & S* & & \\ S_{31} & S_{32} & S* & \\ S_{41} & S_{42} & S_{43} & S* \end{bmatrix}, \text{and} \tag{B2}$$

$$m^{\circ t} = \lfloor m* \quad m* \quad m* \quad m* \rfloor. \tag{B3}$$

Suppose that the four cases are members of family j, then the data are collected in a single vector:

$$y_j^t = \lfloor y_1 \quad y_2 \quad y_3 \quad y_4 \rfloor, \tag{B4}$$

and the contribution of this family to the loglikelihood function is:

$$LogL = \log\Big(|S^{\circ}|^{-1/2}(2\pi)^{-(4p)/2} \times \exp\Big[\Big(-1/2(y_i - m^{\circ})^t S^{\circ -1}(y_j - m^{\circ})\Big)\Big]\Big). \tag{B5}$$

Here we estimate the same parameters as for the model for $S*$ and $m*$ in Eq. (B_1), but in addition we estimate the off-diagonal covariance matrices S_{21} to S_{43}. These off-diagonal matrices accommodate the covariance between family members that are due to their shared genes and shared environment. In practice, the number of off-diagonal matrices is limited by the number of familial relationships. In fact, we limited ourselves to just three off-diagonal matrices: S_{MZ}, the off-diagonal covariance matrix for monozygotic twins, S_{DZ}, the off-diagonal covariance matrix for dizygotic twins, and S_{sibs}, the off-diagonal matrix for full sibs.

For instance, suppose a family j consists of one pair of monozygotic female twins, and 2 male sibs. The model implied covariance matrix for this family is then specified as:

$$S_j^{\circ} = \begin{bmatrix} S_{\mathrm{f}} & & & \\ S_{\mathrm{MZ}} & S_{\mathrm{f}} & & \\ S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{m}} & \\ S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{m}} \end{bmatrix}, \tag{B6}$$

where the subscripts 'MZ' and 'sib' denotes the presence of monozygotic and ordinary sib relations between family members, respectively. In addition, the subscript 'f' denotes the covariance matrix for the females, the subscript 'm' denotes the covariance matrix for the males. That is, the data from the female sibs are used in the composition of the overall covariance matrix for the females, i.e., across all families, and the data from the males sibs are used to compose the overall covariance matrix for the males. For the present example, the mean vector is:

$$m_j^{\circ t} = \lfloor m_{\mathrm{f}} \quad m_{\mathrm{f}} \quad m_{\mathrm{m}} \quad m_{\mathrm{m}} \rfloor. \tag{B7}$$

Again, the data gathered in the female family members are used to establish the overall means for the females, i.e., means across families, while the data from the males are used to establish the overall means for the males.

If the family j consists of one pair of dizygotic male twins, and 1 male and 1 female sib, the matrices $\mathbf{S}_j^{\circ}$ and $\boldsymbol{m}_j^{\circ}$ are:

$$\mathbf{S}_j^{\circ} = \begin{bmatrix} S_{\mathrm{m}} & & & \\ S_{\mathrm{DZ}} & S_{\mathrm{m}} & & \\ S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{m}} & \\ S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{f}} \end{bmatrix}, \text{and} \tag{B8}$$

$$m_j^{\circ t} = \lfloor m_{\mathrm{m}} \quad m_{\mathrm{m}} \quad m_{\mathrm{m}} \quad m_{\mathrm{f}} \rfloor. \tag{B9}$$

If the family j consists of one pair of dizygotic opposite sex twins, and 1 male and 1 female sib, the matrices $\mathbf{S}_j^{\circ}$ and $\boldsymbol{m}_j^{\circ}$ are:

$$\mathbf{S}_j^{\circ} = \begin{bmatrix} S_{\mathrm{m}} & & & \\ S_{\mathrm{sib}} & S_{\mathrm{f}} & & \\ S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{m}} & \\ S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{sib}} & S_{\mathrm{f}} \end{bmatrix}, \text{and} \tag{B10}$$

$$m_j^{\circ t} = \lfloor m_{\mathrm{m}} \quad m_{\mathrm{f}} \quad m_{\mathrm{m}} \quad m_{\mathrm{f}} \rfloor. \tag{B11}$$

In summary, dependence in the data was dealt with by considering families rather than individuals as the unit of analysis. This set-up creates the possibility to estimate the additional covariance matrices between family members, which then function in the model as nuisance parameters. Implementation of this procedure in Mx is discussed in Van der Sluis and Dolan (http://users.fmg.uva.nl/cdolan/).

References

Aluja-Fabregat, A., Colom, R., Abad, F. J., & Juan-Espinosa, M. (2000). Sex differences in general intelligence defined as *g* among young adolescents. *Personality and Individual Differences*, *28*, 813–830.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Azzelini, A. (1996). *Statistical inference based on the likelihood*. London: Chapman & Hall.

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, *20*, 309–328.

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

Boomsma, D. I. (1998). Twin registers in Europe: An overview. *Twin Research*, *1*(1), 34–51.

Born, M. P., & Lynn, R. (1994). Sex differences on the Dutch WISC-R. *Educational Psychology*, *14*(2), 249–254.

Carretta, T. R., & Ree, M. J. (1995). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences*, *19*(2), 149–155.

Carretta, T. R., & Ree, M. J. (1997). Negligible sex differences in the relation of cognitive and psychomotor abilities. *Personality and Individual Differences*, *22*(2), 165–172.

Carroll, J. B. (1997). Psychometrics, intelligence, and public perception. *Intelligence*, *24*(1), 25–52.

Colom, R., Abad, F. J., García, L. F., & Juan-Espinosa, M. (2002). Education, Wechsler's full scale IQ, and *g*. *Intelligence*, *30*, 449–462.

Colom, R., Juan-Espinosa, M., Abad, F. J., & García, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*(1), 57–68.

Colom, R., & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12–18 year olds. *Personality and Individual Differences*, *36*, 75–82.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by *g*. *Intelligence*, *32*, 277–296.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Monkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163–183.

Daniels, M., Devlin, B., & Roeder, K. (1997). The heritability of IQ. *Nature*, *388*, 468–471.

Deary, I. J. (2001). Human intelligence differences: A recent history. *Trends in Cognitive Sciences*, *5*, 127–130.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioural Research*, *35*(1), 21–50.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J., Hessen, D. J., & van der Sluis, S. (submitted for publication). Multi-group covariance and mean structure modeling of the relationship between WAIS-III common factors and gender and educational attainment in Spain. *Intelligence*.

Dolan, C. V., & Hamaker, E. L. (2001). Investigating black–white differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances of Psychological Research*, *vol. 6*. (pp. 31–59). Huntington: Nova Science Publishers.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence*, *32*, 155–173.

Escorial, S., Juan-Espinosa, M., García, L. F., Rebollo, I., & Colom, R. (2003). Does *g* variance change with adulthood? Testing the age de-differentiation hypothesis across sex. *Personality and Individual Differences*, *34*, 1525–1532.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, *44*, 409–420.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179–203.

Gustafsson, J. E. (1992). The relevance of factor analysis for the study of group differences. *Multivariate Behavioral Research*, *27*(2), 239–247.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf–Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment. Theories, tests, and issues*. New York: The Guildford Press.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53–69.

Jensen, A. R. (1981). *Straight talk about mental tests*. London: Methuen.

Jensen, A. R. (1994). Psychometric *g* related to differences in head size. *Personality and Individual Differences*, *17*(5), 597–606.

Jensen, A. R. (1998). *The g factor*. London: Praeger.

Jensen, A. R., & Reynolds, C. R. (1983). Sex differences on the WISC-R. *Personality and Individual Differences*, *4*, 223–226.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.50: User's reference guide*. Chicago: Scientific Software International.

Kimura, D. (1999). *Sex and cognition*. Cambridge, MA: MIT Press.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence*, *14*, 389–433.

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*(3), 299–324.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*, 543–566.

Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences*, *17*(2), 257–271.

Lynn, R. (1998). Sex differences in intelligence: Data from a Scottish standardization of the WAIS-R. *Personality and Individual Differences*, *24*(2), 289–290.

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

Lynn, R., Irwing, P., & Cammock, T. (2001). Sex differences in general knowledge. *Intelligence*, *30*, 27–39.

Lynn, R., Fergusson, D. M., & Horwood, L. J. (2005). Sex differences on the WISC-R in New Zealand. *Personality and Individual Differences*, *39*, 103–114.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*, 110–133.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Millsap, R. E. (1997). The investigation of Spearman's hypothesis and the failure to understand factor analysis. *Cahiers de Psychologie Cognitive*, *16*, 750–757.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling* (6th edition). VCU Box 900126, Richmond, VA 23298: Department of Psychiatry.

Nyborg, H. (2003). Sex differences in *g*. In H. Nyborg (Ed.), *The scientific study of general intelligence* (pp. 187–227). Amsterdam: Elsevier Science Ltd.

Pennington, B. F., Filipek, P. A., Lefly, D., Chhabildas, N. A., Kennedy, D. N., Simon, J. H., et al. (2000). A twin MRI study of size variation in the human brain. *Journal of Cognitive Neuroscience*, *12*(1), 223–232.

Rushton, J. P. (1992). Cranial capacity related to sex, rank, and race in a stratified random sample of 6325 U.S. military personnel. *Intelligence*, *16*, 401–413.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Schrijvers, C. T. M., Stronks, K., Van de Mheen, H., & Mackenbach, J. P. (1999). Explaining educational differences in mortality: The role of behavioral and material factors. *American Journal of Public Health*, *89*, 535–540.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

Stronks, K., Van de Mheen, H., & Mackenbach, J. P. (1997). The interrelationship between income, health and employment status. *International Journal of Epidemiology*, *26*, 592–600.

Undheim, J. O., & Gustafsson, J. E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, *22*, 149–171.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and considerations of critical variables. *Psychological Bulletin*, *117*(2), 250–270.

WAIS-III. (1997). *Dutch version. Manual*. Lisse: Swets and Zeitlinger.

Wechsler, D. (1998). *WAIS-IIIR manual*. New York: The Psychological Corporation.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509–537.

Wickett, J. C., Vernon, P. A., & Lee, D. H. (2000). Relationships between factors of intelligence and brain volume. *Personality and Individual Differences*, *29*, 1095–1122.

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.