Available online at www.sciencedirect.com
ScienceDirect
Intelligence 36 (2008) 48–67
INTELLIGENCE

Sex differences on the WISC-R in Belgium and The Netherlands

Sophie van der Sluis [a,*], Catherine Derom [b], Evert Thiery [c], Meike Bartels [a], Tinca J.C. Polderman [a,d], F.C. Verhulst [d], Nele Jacobs [c], Sofie van Gestel [c], Eco J.C. de Geus [a], Conor V. Dolan [e], Dorret I. Boomsma [a], Danielle Posthuma [a]

[a] *Department of Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands*
[b] *Department of Human Genetics, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*
[c] *Association for Scientific Research in Multiple Births, B-9070 Destelbergen, Belgium*
[d] *Department of Child and Adolescent Psychiatry, Erasmus MC- Sophia, Dr. Molewaterplein 60, 3015 GJ, Rotterdam, The Netherlands*
[e] *Department of Psychology, FMG, University of Amsterdam, Roeterstraat 15, 1018 WB, Amsterdam, The Netherlands*

Received 6 May 2006; received in revised form 15 January 2007; accepted 17 January 2007
Available online 22 February 2007

Abstract

Sex differences on the Dutch WISC-R were examined in Dutch children (350 boys, 387 girls, age 11–13 years) and Belgian children (370 boys, 391 girls, age 9.5–13 years). Multi-group covariance and means structure analysis was used to establish whether the WISC-R was measurement invariant across sex, and whether sex differences on the level of the subtests were indicative of sex differences in general intelligence (*g*). In both samples, girls outperformed boys on the subtest Coding, while boys outperformed girls on the subtests Information and Arithmetic. The sex differences in the means of these three subtests could not be accounted for by the first-order factors Verbal, Performance, and Memory. Measurement invariance with respect to sex was however established for the remaining 9 subtest. Based on these subtests, no significant sex differences were observed in the means of the first-order factors, or the second-order *g*-factor. In conclusion, the cognitive differences between boys and girls concern subtest-specific abilities, and these sizeable differences are not attributable to differences in first-order factors, or the second-order factor *g*.

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

1. Introduction

Sex differences on the WISC-R have been studied in the WISC-R standardization samples of the USA, Scotland, The Netherlands, and China, and in data from Mauritius, New Zealand, and Belgium (e.g., Born & Lynn, 1994; Dai & Lynn, 1994; Grégoire, 2000; Jensen & Reynolds, 1983; Lynn & Mulhern, 1991; Lynn, Riane, Venables, Mednick, & Irwing, 2005). The results are largely comparable across countries. Consistently, large differences favoring girls are reported regarding the subtest Coding (effect sizes about .5), and large differences favoring boys are reported regarding the subtest Information (effect sizes about .35). In addition, girls sometimes outperform boys on the subtest Digit Span, but these differences are usually small and statistically insignificant. Boys score slightly higher than girls on all other subtests, and even though these differences are sometimes statistically significant,

* Corresponding author.
E-mail address: s.van.der.sluis@psy.vu.nl (S. van der Sluis).

0160-2896/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2007.01.003

the differences are often small, with effect sizes ranging between .00 and .20.

In all these studies, WISC-R subtest scores and factor scores have been compared directly between boys and girls. Yet it has never been established whether the factor structure of the WISC-R is actually comparable or 'measurement invariant' across sex (see below). The interpretation of group differences in subtest- or factors scores may be complicated greatly if the underlying factor structure differs between the groups. That is, if a test battery does not measure the same construct(s) in different groups, then group differences in test scores representing first or higher order factors are difficult to interpret. The aim of the present study is to find out whether the WISC-R is measurement invariant across sex in children before comparing subtest and factor scores between boys and girls.

The factor structure underlying the WISC-R has been studied in clinical and non-clinical samples (e.g., Anderson & Dixon, 1995; Burton et al., 2001; Donders, 1993; Huberty, 1987; Kush et al., 2001; Meesters, van Gastel, Ghys, & Merckelbach, 1998; Wright & Dappen, 1982). Principal component analyses (PCA, e.g., Born & Lynn, 1994; Lynn & Mulhern, 1991; Rushton & Jensen, 2003), exploratory factor analyses (EFA, e.g., Dolan, 2000; Dolan & Hamaker, 2001; Kush et al., 2001), and confirmatory factor analyses (CFA, e.g., Burton et al., 2001; Dolan, 2000; Dolan & Hamaker, 2001; Keith, 1997; Kush et al., 2001; Oh, Glutting, Watkins, Youngstrom, & McDermott, 2004) have yielded either a two factor ('Verbal' and 'Performance'), or a three factor solution ('Verbal', 'Performance', and 'Memory', also known as 'Freedom from distractibility'). In these models, general intelligence ('*g*') was either operationalized as the first principal component (PCA), or as a second-order factor (CFA).

Given the assumption that these latent factors underlie the performance on the level of the subtests, one question of interest is whether the observed sex differences at the level of the subtests are a function of differences in *g*, or of differences on the level of the broad primary factors of intelligence (e.g., Verbal intelligence, Performance intelligence and Memory). However, it may also be the case that the subtest differences are not attributable to common factor differences, but rather are a manifestation of differences in the specific ability that the subtest taps.

If boys and girls differ with respect to the mean on a given subtest, and this difference cannot be explained by the mean differences on the latent factor, which is supposed to underlie performance on the subtest, then the subtest may be viewed as biased with respect to sex. The term bias does not imply that the observed mean difference is not real, rather the term, as used here, implies that the mean difference on the subtest is greater or smaller than that expected on the basis of the latent factor mean difference. According to this definition, the term bias refers to the subtest as an indicator of the common factor, which the subtest is supposed to measure. For example, it has been established that the Information subtest of the WAIS is biased with respect to sex. Specifically, the male advantage on this subtest, which is supposed to measure general knowledge, is too large to be accounted for by the common factor Verbal Comprehension (e.g., Dolan et al., 2006; Van der Sluis et al., 2006). The difference is not indicative of a difference with respect to Verbal Comprehension. However, it may well be indicative of a true male advantage in general knowledge.

Establishing the exact nature of an observed (subtest) mean difference is important in the light of theories, in which sex differences are attributed to latent mean differences (e.g., a difference in Verbal Comprehension, or a difference in *g*). In previous studies aimed at identifying the source(s) of the sex differences, PCA was mostly used to investigate sex differences on the factors underlying intelligence. Sex differences were evaluated by calculating weighted linear combination of the subtests means, where the subtests' factor loadings served as weights (e.g., Born & Lynn, 1994; Jensen & Reynolds, 1983; Lynn, Fergusson, & Horwood, 2005; Lynn & Mulhern, 1991; Lynn, Riane, et al., 2005). The general finding of these studies is that boys score higher on the Verbal and Performance factors, while girls score higher on the Memory factor. With respect to general intelligence, operationalized as the first principal component, boys usually score higher than girls, but effect sizes are often small (about .10), and the difference is not always statistically significant. When expressed on the conventional IQ-scale with a mean of 100 and standard deviation of 15, these sex differences range from 1 to 6 IQ points (e.g., Lynn, Fergusson, et al., 2005; Lynn, Riane, et al., 2005). All these results are however based on samples with a broad age-range (6–16 years), and it remains to be seen whether the factor structure of the WISC-R, and the effects reported for the (factor) means, are stable across age.

One obvious problem concerning this PCA-based method of studying sex differences is that sex differences on the level of the weighted means of the observed subtest scores may be due to one or just a few of many subtests. For example, boys may outperform girls on the Verbal factor only because they outperformed girls on the subtest Information, while their performance on the other verbal subtests may even be inferior. In that case, it

is more accurate to conclude that boys outperform girls with respect to a specific cognitive ability, i.e., general knowledge, rather than suggesting that boys have higher Verbal intelligence. Stated succinctly, this method does not explicitly address the structure of the observed mean differences. The use of PCA to study group differences is however characterized by several other disadvantages. Because PCA, in contrast to EFA or CFA, is based on a data transformation rather than an explicit statistical model, it does not generally include statistical testing or explicit model fitting. As a consequence, goodness of fit is not evaluated, i.e., the question of whether the model provides a reasonable description of the structure of the data is not addressed. In addition, an explicit statistical procedure to test whether the factor structure underlying the given test battery is comparable across groups is not conducted. The comparability of the factor structure across groups is however of utmost importance if one wishes to make meaningful comparisons between the subtest scores or factor scores of different groups. Furthermore, within the context of PCA, competing hypotheses are not compared statistically (e.g., are sex differences present on the level of the primary factors of intelligence or rather on the level of the observed subtests only?). Finally, PCA is not suited for modeling measurement error in the subtest scores, which is sure to exist.

An alternative method for testing group differences within the context of factor models is multi-group covariance and means structure analysis (MG-CMSA; Sörbom, 1974; Little, 1997; Widaman & Reise, 1997). This method provides a comprehensive, model-based means to investigate the main sources of group difference. MG-CMSA allows one to evaluate and compare the fit of alternative models, which correspond with competing hypotheses. The advantages of MG-CMSA have been studied and discussed in detail, and MG-CMSA has repeatedly been shown to be superior to other methods used to study group difference (e.g., method of correlated vectors, the Schmidt–Leiman procedure, PCA) with regard to, among things, its flexibility and the facility to test (competing) hypotheses explicitly (see e.g., e.g., Dolan, 2000; Dolan & Hamaker, 2001; Dolan, Roorda, & Wicherts, 2004; Lubke, Dolan, & Kelderman, 2001; Lubke, Dolan, Kelderman, & Mellenbergh, 2003; Millsap, 1997). Previously, MG-CMSA was used to study ethnic group difference in intelligence (e.g., Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Gustafsson, 1992), the Flynn-effect (Wicherts et al., 2004), and sex differences on the WAIS (Dolan et al., 2006; Van der Sluis et al., 2006).

In the present study, we used MG-CMSA to investigate sex differences on the Dutch WISC-R in Dutch and Belgian children of limited age-range (9–13 years old). Below, we outline the MG-CMSA modeling procedure that we used to investigate the sources of sex differences on the WISC-R. Both first- and second-order factor models are fitted, with the second-order factor representing *g*. The results are presented for Dutch and Belgian subjects separately, and are discussed in the light of previous studies.

2. Method

2.1. Subjects

2.1.1. Dutch sample

The Dutch data constitute a combination of two datasets: data that were previously used in a study of the genetic and environmental contributions to the development of individual differences in intelligence (Bartels, Rietveld, van Baal, & Boomsma, 2002), and data that were used to establish the extent to which the phenotypic correlation between working memory speed and capacity is of genetic origin (Polderman et al., 2006). All Dutch subjects were recruited from the young Netherlands Twin Register (NTR, Boomsma, 1998; Boomsma et al., 2002; Bartels, Beijsterveldt, Stroet, Hudziak, & Boomsma, in press). Since 1986, the majority of parents with multiple births in The Netherlands receive a brochure and a registration form from the NTR. Registration is voluntary, and about 40% of the parents register their twins with the NTR. Information from questionnaires, blood group, and DNA polymorphisms (genetic markers) was used to assign zygosity to same-sex twins (Rietveld et al., 2000).

For this study, data were available from 368 twin pairs (77 monozygotic male pairs, 100 monozygotic female pairs, 67 dizygotic male pairs, 62 dizygotic female pairs, and 62 opposite sex twins), and, due to missingness, one single twin. As in most of the twin studies, the percentage of MZ twins in this sample (48%) is somewhat higher than in the overall population (~ 33%) due to self-selection bias.

The sample included 350 boys and 387 girls (737 subjects in total). For all twins in this study, level of parental occupation was assessed at age 10 of the twins. Occupational level was rated on a 5-point scale, ranging from manual labor to academic employment. Paternal occupational level was used, or maternal occupational level in case paternal information was not available. In comparison to the Dutch population (Centraal Bureau voor de Statistiek, 2002), the level of occupation of the

parents of the twins participating in this study was somewhat higher: the percentages observed in the present study and the Dutch population are: 1% and 6% (manual), 15% and 26% (lower), 42% and 40% (middle), 30% and 19% (higher), and 11% and 9% (academic). Paternal and maternal educational level did not differ between the boys and girls in this sample ($z = -.19$, ns, and $z = -.20$, ns, respectively).

With the exception of one twin pair aged 10.9 years old, the age of all subjects ranged between 11.9 and 12.9 years at the time of testing. The youngest twin pair at 10.9 years old was not removed from the sample as it did not constitute an outlier in any aspect. Sex differences with respect to age were absent ($t(735) < 1$, ns).

2.1.2. Belgian sample

The Belgian subjects were recruited from the East Flanders Prospective Twins Survey (EFPTS), a population-based register of twins in the province of East Flanders, Belgium (Derom et al., 2002; Loos, Derom, Vlietinck, & Derom, 1998). Since 1964, EFPTS collects information on the mother, the placenta and the child of 98% of all multiples born in the province. Zygosity of all twins was determined through sequential analysis based on sex, foetal membranes, umbilical cord blood groups (ABO, Rh, CcDEe, Mnss, Duffy, Kell), placental alkaline phosphatase and, since 1982, DNA fingerprints. Unlike-sex twins and same-sex twins with at least one different genetic marker were classified as DZ; monochorionic twins were classified as MZ. For all same-sex dichorionic twins with the same genetic markers a probability of monozygosity was calculated.

All subjects, whose data are used in the present study, participated in an ongoing study on cognitive ability in twins aged 7.5 to 15 years old. This sample was shown to be representative for gender, birth weight, and gestational age. As in most of the twin studies, the MZ twins were slightly over represented (42%) due to self-selection biases. Comparison of the 663 twins with known IQ scores with the twins who refused to participate in the study ($n = 204$) revealed that, in the non-participating group, parents with a lower educational level and twins who attend special schools tend to be over represented. Part of the data (only complete same-sexed twin pairs with known IQ scores) was previously used to study the effect of chorion-type on the estimation of the heritability of intelligence (Jacobs et al., 2001).

The present dataset comprises a subsample (age-range between 9.5 and 13 years at time of measurement) of the above-mentioned study. The sample consisted of 370 boys and 391 girls (761 subjects in total). Data were available from 374 complete twin pairs (83 monozygotic male pairs, 76 monozygotic female pairs, 44 dizygotic male pairs, 63 dizygotic female pairs, and 108 opposite sex twin pairs), and 13 single twins. Sex differences with respect to age were absent ($t(759) = 1.73$, ns).

2.2. Tests

In both Dutch and Belgian samples, psychometric IQ was measured with the following 12 subtests of the Dutch WISC-R (Van Haasen et al., 1986): Information (INF), Similarities (SIM), Arithmetic (AR), Vocabulary (VOC), Comprehension (COMP), Picture Completion (PC), Picture Arrangement (PA), Block Design (BD), Object Assembly (OA), Mazes (MA), Coding (CO), and Digit span (DS). In the Belgian data missingness was absent. The Dutch data, in contrast, included systematic missing data. Specifically, for reasons of efficiency, only 6 out of 12 subtests (namely SIM, AR, VOC, BD, OA, and DS) were administered to the 354 Dutch subjects (165 boys, 189 girls) who previously participated in the study by Polderman et al. (2006). Some additional missingness occurred due to procedural errors, but this percentage was very small (.2%).

Due to this systematic omission of subtests in part of the Dutch sample, missingness can not be considered completely at random (MCAR, e.g., Schafer & Graham, 2002) in the total Dutch sample ($p < .01$ for Little's MCAR test performed across families and for boys and girls separately). In the following exploratory and confirmatory factor analyses, raw data Maximum Likelihood estimation was employed to accommodate the missingness, and use all available data. Raw data ML estimation has been found to provide better parameter estimates than conventional methods, such as listwise or pairwise deletion and mean imputation, even if data are not missing (completely) at random (e.g., Tomarken & Waller, 2005).

2.3. Statistical analyses

2.3.1. Measurement invariance and model fitting strategies

To study sex differences in the means and covariances within the common factor model, multi-group confirmatory covariance and means structure analysis (MG-CMSA) was used. Before sex differences with respect to the latent common factors can be examined, we first need to establish whether the WISC-R is measurement invariant with respect to sex. Measurement invariance with regard to sex implies that the distribution of the observed scores on a subtest (y_i) given a fixed level of the

latent factor (η), depends on the score on the latent factor η, and not sex, i.e., $f[y_i|\eta,\text{sex}]=f[y_i|\eta]$ (Mellenbergh, 1989). Given normally distributed data, measurement invariance can be defined in terms of the means and the variances of the y_i given η. With respect to the means, measurement invariance implies that the expected value of subjects *i* on subtest *y* depends only on the latent factor score η, and not on sex, i.e., $E[y_i|\eta,\text{sex}]=E[y_i|\eta]$. Within the common factor model, to establish measurement invariance one needs to establish whether the relation between the observed subtest scores and the underlying latent factors is the same in boys and girls (Meredith, 1993). Measurement invariance can be established through the imposition of a series of specific constraints on the model parameters over groups, i.e., across sex in this case (Meredith, 1993). First of all, the subtests should load on the same factors in both boys and girls, i.e., the measurement model should be the same in both sexes (also called configural invariance). Subsequently, the function relating the observed subtest scores to the latent factors can be considered identical for boys and girls if the following parameters can, to reasonable approximation, be considered equal across sex: a) the factor loadings of the observed subtests on the latent factors, b) the intercepts (note that the factor means are allowed to differ across groups), and c) the residual variances, i.e., the variance in the observed subtest scores that is not explained by the latent common factors. If these constraints prove tenable, the WISC-R may be considered to be measurement invariant with respect to sex, and in that case, individual differences and group differences on the level of the observed subtests can be interpreted in terms of differences on the common factors.

The model fitting strategy that follows from the above described equality constraints is described in detail in Van der Sluis et al. (2006). Below we give an overview of this model fitting procedure, and we refer to Appendix A of Van der Sluis et al. (2006) for a description of this procedure in matrix notation.

2.3.2. First-order factor models

First, we fitted the least constrained model, model F_1, that tests for configural invariance (Horn & McArdle, 1992; Widaman & Reise, 1997). Configural invariance implies that the configuration of factor loadings (and correlated residuals, if any) is the same across sex, but the exact values of these parameters are allowed to differ across groups. In this model, the observed means of the 12 subtests are estimated freely in boys and girls, i.e., we do not yet introduce a constrained model for the mean structure. Note that in model F_1, we fixed the variances of the latent factors to 1 in both boys and girls. This is a standard identifying constraint in factor analysis (e.g., see Bollen, 1989).

Subsequently, we tested for metric invariance (Horn & McArdle, 1992; Widaman & Reise, 1997) by constraining the factor loadings to be identical across sex. We denote this model F_2. Identical factor loadings are a prerequisite for a meaningful comparison between boys and girls with respect to the latent common factors, i.e., if the factor loadings of the subtests on the latent factors are not identical across sex, we cannot be sure that the latent factors are identical, and thus comparable, across sex. If the constraints introduced in model F_2 do not result in a significant deterioration of the model fit compared to model F_1, metric invariance is considered tenable. Note that these equality constraints on the factor loadings render fixation of the factorial variance in both group superfluous, so in model F_2, the factor variances remain fixed to 1 in the boys, but are estimated freely in the girls.

Next, we test for strong factorial invariance (Horn & McArdle, 1992; Meredith, 1993; Widaman & Reise, 1997) by introducing a restrictive structure for the means. We denote this model F_3. In model F_3, the intercepts in the regression of the observed variables on the common factors are constrained to be equal in boys and girls, while the means of the factors are estimated. We thus introduce a constrained model for the means structure. Note that for reasons of identification, it is not possible to estimate the factor means in both groups (Sörbom, 1974). We chose to fix the factor means to zero in the boys, and estimated freely in the girls. Modeled as such, the boys function as a reference group, and the factor means in the female group are calculated as deviations from the factor means of the boys. If the fit of model F_3 is not significantly worse than the fit of model F_2, the assumption that the expected values of the observed subtest scores depend not on sex but only on the latent factor scores, is considered tenable, i.e., $E[y_i|\eta,\text{sex}]=E[y_i|\eta]$. Model F_3 thus embodies the test whether the latent common factors can account for the observed mean differences between boys and girls on the level of the subtests. Boys and girls can be compared meaningfully with respect to their first-order factors means only if model F_3 holds. If model F_3 is not tenable, one or more of the mean differences between boys and girls on the level of the observed subtest scores cannot be accounted for by the first-order factors. As explained above, subtests are considered biased with respect to sex if observed difference on these tests cannot be attributed to differences on the level of the primary factors of intelligence.

We next test for strict factorial invariance (Horn & McArdle, 1992;Meredith, 1993; Widaman & Reise, 1997) by constraining the residual variances to be equal across sex. We denote this model F_4. If model F_4 is

tenable, in comparison to model F_3, we conclude that all differences between boys and girls with respect to the means and the covariance structure can be accounted for by sex differences in the first-order factors. Note that the tenability of model F_4 is not a prerequisite for the comparability of boys and girls with respect to the observed means, or with respect to the means of the first- and second-order latent factors. To this end, F_3 suffices.

2.3.3. Second-order factors

A second-order factor (model S_1), i.e., the model including general intelligence *g* as a second-order factor, was introduced in either model F_3 or F_4, depending on the tenability of the constraints introduced in model F_4. Depending on the number of first-order factors, this hierarchical factor model is either equivalent to the first-order factor model (in the case of 3 first-order factors), or it tests whether all relations between the first-order factors can be explained by 1 second-order factor (in the case of 4 or more first-order factors). At this point, the second-order factor loadings, i.e., the loadings of the first-order factors on the second-order factor, are allowed to differ across sex, and the means of the second-order factors are fixed to zero in both groups, while the first-order factor means are fixed to zero in boys, and freely estimated in girls (as in models F_3 and F_4). For reasons of identification, the variance of the second-order factor is fixed to 1 in both groups.

In model S_2, the second-order factor loadings are constrained to be equal across sex. Like in model F_2, these equality constraints on the factor loadings allow one to freely estimate the variance of the second-order factor in one of the groups (in our case the girls). With model S_2 we thus test whether the factor loadings of the first-order factors on the second-order factor are equal in boys and girls.

In model S_3, the first-order factor means are constrained to be equal across sex. Given our present parameterization, this involves fixing the first-order factor means differences to zero in the girls. The second-order factor mean is then fixed to zero in the boys, and estimated freely in the girls (analogous to model F_3). In this model, the mean difference between boys and girls are described entirely in terms of mean differences on the second-order factor, i.e., in *g*. If model S_3 is tenable (in comparison to model S_2), we conclude that the mean differences between boys and girls on the level of the observed subtests can be accounted for completely by differences in *g*. If model S_3 does not fit the data, we conclude that the differences between boys and girls at the level of the first-order factors are not (or not completely) attributable to difference between boys and girls in *g*.

In the final model, model S_4, we constrain the second-order factor means to be equal across sex (i.e., we fix the second-order factor mean difference to zero). If model S_4 fits as well as model S_3, we conclude that boys and female do not differ with respect to *g*. A significant deterioration of the fit as a result of this constraint indicates the presence of sex differences in *g*.

2.3.4. Estimation and model fit

Both the Dutch and Belgian data were gathered within families. The focus of this paper, however, is on gender differences, and not on the correlations among family members, which are undoubtedly present as cognitive abilities are known to be quite heritable (e.g., Bartels et al., 2002; Daniels, Devlin, & Roeder, 1997; Posthuma et al., 2002). Treating within-family data as if they are independently distributed observations results in incorrect standard errors and incorrect χ^2 goodness of fit values, while the point estimations of parameter estimates remain unbiased (e.g., Rebollo, de Moor, Dolan & Boomsma, 2006). All factor analytic analyses were therefore performed in Mplus, version 4 (Muthén & Muthén, 2005), which computes corrected standard errors and Satorra–Bentler scaled χ^2-tests, taking into account the dependence of observations. Competing hypotheses, represented by different nested models (where the nested model is the more restricted model), can be compared through a weighted χ^2-difference test developed especially for the comparison of the Satorra–Bentler scaled χ^2s (Satorra, 2000). The more restricted model is accepted as the preferred model, if its fit is not significantly worse than the fit of the less restrictive model, i.e., if the χ^2-difference test (henceforth the χ^2_{diff}) is not significant. Below, we will not report scaled χ^2-values for each model separately, as these are not informative, rather we report weighted χ^2_{diff} tests for the comparison between competing models. Given the large sample sizes, and the number of tests that were required to compare all ensuing models, we chose an α of .01.

To evaluate the fit of the ensuing models to the data, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI) were used (e.g., Bentler, 1990; Bollen & Long, 1993; Jöreskog, 1993; Schermelleh-Engel, Moosbrugger, & Müller, 2003). The RMSEA is a measure of the error of approximation of the covariance and mean structures as implied by the specified model to the covariance and mean structures in the population. As a measure of approximation-discrepancy per degree of freedom, this fit index favors more parsimonious models. Generally, good fitting models are thought to have RMSEA $< .05$, although simulation studies by Hu and Bentler (1999) showed

that a cut-off criterion of .06 can be used as well. Here we adopt the following rule of thumb: a RSMEA of .05 or less indicates good approximation, RMSEA between .05 and .08 indicates reasonable approximation, and RMSEA greater than .08 indicates poor approximation (Browne & Cudeck, 1993; Schermelleh-Engel et al., 2003). The CFI is based on the comparison of the fit of the target model (i.e., the user-specified model) with the fit of the independence model (i.e., a model in which all variables are modeled as unrelated). Like the RMSEA, the CFI favors more parsimonious models. CFI ranges from zero to 1.00, and values >.90 or .95 are usually taken as indicative of adequate model fit (e.g., Hu & Bentler, 1999; Schermelleh-Engel et al., 2003). When testing for measurement invariance, the scaled χ^2 statistic was used to compare the fit of the competing models, while the RMSEA and the CFI were used only to check that the general fit of the ensuing models was still acceptable. In addition, modification indices were used to detect local misspecifications in the models. The modification index of a constrained parameter (i.e., fixed to a given value or subject to an equality constraint) expresses the expected drop in overall χ^2, if the constraint on the parameter is relaxed.

3. Results

All analyses were performed on the standardized subtest scores, which have a mean of 10 and SD of 3 in the population.

3.1. Preliminary analyses

Table 1 contains means and standard deviations of all 12 standardized subtest scores, reported separately for Dutch and Belgian boys and girls. As a measure of effect size, Cohen's *d* is also reported, which is calculated as the difference between the mean of the boys and the girls ($\mu_{boys} - \mu_{girls}$) divided by the pooled standard deviation, so that positive (negative) *d*'s denote male (female) advantage. Most effect sizes were small (<|.3|), and medium effect sizes (between |.3| and |.6|) were only observed with respect to INF and AR (favoring boys) and CO (favoring girls).

It is possible that the differences between boys and girls are indicative of differences between families in, for example, socioeconomic status (SES), rather than of genuine sex differences. It is impossible to measure all variables on which families might differ, but comparing boys and girls who grew up in the same family environment provides a powerful check of the possible influence of family background. Opposite sex twin pairs are therefore of special interest. If the sex differences that are observed across families remain significant within families, then these differences are more likely to represent real differences between boys and girls. However, if these between family differences diminish, or even disappear, within families, the between family differences are more likely to relate to environmental differences. (Note that the opposite twin design does not imply perfect matching of boys and girls; e.g.,

Table 1
Means (M) and standard deviations (SD) for the Dutch and Belgian boys and girls on the 12 WISC-R subtests

	Netherlands							Belgium						
	Boys			Girls				Boys			Girls			
	M	SD	*N*	M	SD	*N*	*d*	M	SD	*N*	M	SD	*N*	*d*
INF	10.33	2.47	185	8.98	2.75	198	.52	9.45	2.81	370	8.46	2.56	391	.37
SIM	9.71	3.03	349	9.56	2.96	385	.05	10.92	2.91	370	10.90	3.15	391	.01
AR	11.03	2.91	348	10.14	2.83	387	.31	10.17	2.90	370	9.63	2.92	391	.19
VOC	9.27	2.31	349	8.66	2.29	387	.27	11.11	2.74	370	10.91	2.66	391	.07
COMP	9.76	2.25	185	9.36	2.34	198	.17	11.34	2.79	370	11.49	2.97	391	−.05
PC	11.41	2.72	185	11.01	2.51	198	.15	9.91	2.86	370	9.80	2.81	391	.04
PA	10.05	2.94	185	9.52	2.87	198	.18	10.15	2.90	370	9.54	3.10	391	.20
BD	10.35	3.02	350	10.15	2.84	387	.07	10.31	2.94	370	9.91	3.10	391	.13
OA	8.68	2.84	350	8.67	2.88	384	.00	9.53	3.17	370	9.03	3.19	391	.16
CO	10.93	2.61	183	12.35	2.75	196	−.53	9.67	2.73	370	11.21	3.12	391	−.53
MA	11.34	2.82	185	10.93	2.68	198	.15	10.53	2.96	370	10.03	2.99	391	.17
DS	10.25	2.70	348	10.34	2.46	386	−.03	10.22	2.90	370	10.58	2.91	391	−.12

Scores are standardized scores (in norm sample, $M=10$, $SD=3$).
Note. *d* is Cohen's measure of effect size *d*, defined as ($M_{boys} - M_{girls}/\sigma_{pooled}$).
INF=Information, SIM=Similarities, AR=Arithmetic, VOC=Vocabulary, COMP=Comprehension, PC=Picture Completion, PA=Picture Arrangement, BD=Block Design, OA=Object Assembly, MA=Mazes, CO=Coding, DS=Digit span.

systematic differences may exist in the way that boys and girls are treated). To more closely examine the mean differences, paired *t*-tests were performed on the data of the Dutch and Belgian opposite sex twin pairs (Table 2).

In both the Dutch and Belgian samples, boys scored significantly higher on INF than their female sibling, while girls scored significantly higher on CO than their male sibling. These results are consistent with the effect sizes in Table 1, and likely to represent genuine differences between boys and girls. In addition, Dutch boys scored higher on AR, VOC and COMP than their female siblings. These latter results are consistent with the intermediate effect sizes for the Dutch sample as presented in Table 1. In sum, the results of the paired *t*-tests correspond to the differences observed between boys and girls in the total sample, and the sex differences are therefore not likely to be the result of between family differences in factors like SES.

The mean differences between boys and girls on the level of the observed subtest scores, and the relation with the underlying primary factors of intelligence are further examined using MG-CMSA.

3.2. Exploratory factor analyses

Because the reported patterns of factor loadings vary across studies, exploratory factor analyses (EFA) were carried out first to establish the pattern of factor loadings in Dutch and Belgian boys and girls separately. The exploratory factor solution was followed by an oblique rotation (Promax; see Lawley & Maxwell, 1971), using normal theory maximum likelihood estimation (ML).

In both Dutch and Belgian samples, solutions with one factor or with two correlated factors were inadequate, while the solution with three correlated factors proved reasonable in terms of goodness of fit and interpretability. Table 3 contains the loadings of the 12 subtests on the three correlated common factors reported separately for Dutch and Belgian boys and girls. Factor loadings in bold print were considered substantial in all subsamples. This empirically established pattern of factor loadings is largely similar to factor solutions

Table 2
Paired *t*-tests for Dutch and Belgian opposite sex twins

	Netherlands (N_{pairs}=62)			Belgium (N_{pairs}=108)		
	t	*df*	*p*	*t*	*df*	*p*
INF	6.53	34	.00	3.70	107	.00
SIM	1.18	61	.24	−.55	107	.59
AR	2.47	61	.02	1.24	107	.22
VOC	2.29	61	.03	.29	107	.77
COMP	2.19	34	.04	−.10	107	.92
PC	−.77	34	.44	.33	107	.74
PA	1.70	34	.10	.52	107	.60
BD	.00	61	1.00	.69	107	.49
OA	.51	60	.61	.72	107	.47
CO	−2.99	33	.01	−4.23	107	.00
MA	.13	34	.90	1.04	107	.30
DS	.45	60	.65	−.15	107	.88

Note. Positive *t*-values indicate male advantage; negative *t*-values indicate female advantage.
INF = Information, SIM = Similarities, AR = Arithmetic, VOC = Vocabulary, COMP = Comprehension, PC = Picture Completion, PA = Picture Arrangement, BD = Block Design, OA = Object Assembly, MA = Mazes, CO = Coding, DS = Digit span.

Table 3
Results exploratory factor analyses, separately for Dutch and Belgian boys and girls

	Netherlands					
	Boys (*N*=350)			Girls (*N*=387)		
	V	P	M	V	P	M
INF	**0.656**	0.083	0.055	**0.761**	−0.021	0.120
SIM	**0.533**	0.044	0.174	**0.700**	0.012	0.008
AR	**0.305**	−0.045	**0.611**	**0.254**	−0.095	**0.670**
VOC	**0.945**	−0.156	−0.010	**0.797**	0.021	−0.001
COM	**0.517**	0.119	0.075	**0.749**	−0.064	−0.021
PC	0.192	**0.274**	−0.049	0.155	**0.401**	0.003
PA	0.257	**0.318**	−0.088	0.301	**0.351**	0.073
BP	−0.169	**0.617**	**0.544**	0.026	**0.593**	**0.213**
OA	−0.098	**0.783**	0.079	−0.100	**0.876**	−0.069
CO	−0.083	0.041	**0.592**	−0.096	0.102	**0.562**
MA	−0.052	0.046	**0.502**	0.044	0.098	**0.257**
DS	0.239	−0.036	**0.405**	−0.083	−0.044	**0.653**
	Belgium					
	Boys (*N*=370)			Girls (*N*=391)		
	V	P	M	V	P	M
INF	**0.506**	0.080	0.289	**0.701**	0.017	0.113
SIM	**0.694**	0.034	−0.010	**0.556**	0.046	0.110
AR	**0.273**	0.010	**0.510**	**0.254**	−0.041	**0.558**
VOC	**0.672**	−0.012	0.196	**0.891**	−0.009	−0.008
COMP	**0.690**	−0.002	−0.001	**0.585**	0.072	0.079
PC	0.105	**0.471**	−0.072	0.040	**0.489**	0.100
PA	0.213	**0.366**	0.054	0.229	**0.425**	0.027
BD	−0.151	**0.458**	**0.589**	−0.026	**0.581**	**0.313**
OA	−0.019	**0.721**	0.097	0.009	**0.901**	−0.101
CO	0.016	0.012	**0.501**	−0.014	0.112	**0.424**
MA	−0.009	0.227	**0.259**	−0.122	0.211	**0.333**
DIG	0.219	−0.113	**0.478**	0.023	−0.119	**0.738**

INF = Information, SIM = Similarities, AR = Arithmetic, VOC = Vocabulary, COMP = Comprehension, PC = Picture Completion, PA = Picture Arrangement, BD = Block Design, OA = Object Assembly, MA = Mazes, CO = Coding, DS = Digit span, V = Verbal factor, P = Performance factor, M = Memory factor.

Table 4
Fit statistics for the Dutch models

		CFI	RMSEA	χ^2_{diff}
F_1	Configural invariance	.96	.04	
F_{1a}	Configural invariance+residuals OA and BD correlated	.98	.03	F_{1a} vs. F_1: $\chi^2_{diff}(2)=26.84$, $p<.001$
F_2	Metric invariance	.98	.03	F_2 vs. F_{1a}: $\chi^2_{diff}(11)=9.43$, ns
F_3	Strong factorial invariance	.95	.05	F_3 vs. F_2: $\chi^2_{diff}(9)=65.66$, $p<.001$
F_{3a}	Strong factorial invariance, bar INF, AR and CO	.98	.03	F_{3a} vs. F_2: $\chi^2_{diff}(6)=10.75$, ns
F_4	Strict factorial invariance	.98	.03	F_4 vs. F_{3a}: $\chi^2_{diff}(13)=12.80$, ns
S_1	Introduction 2nd order factor	.98	.03	S_1 is identical to F_4
S_2	Metric invariance 2nd order factor	.98	.03	S_2 vs. S_1: $\chi^2_{diff}(2)=5.34$, ns
S_3	Strong factorial invariance 2nd order factor	.98	.03	S_3 vs. S_2: $\chi^2_{diff}(2)=1.79$, ns
S_4	Strict factorial invariance 2nd order factor	.97	.03	S_4 vs. S_3: $\chi^2_{diff}(1)=4.20$, ns ($p=.04$)

reported in previous papers (e.g., Burton et al., 2001; Dolan, 2000; Dolan & Hamaker, 2001; Keith, 1997; Kush et al., 2001; Meesters et al., 1998; Oh et al., 2004), with the first factor representing the Verbal factor, the second factor the Performance factor, and the third factor the Memory factor (also known as Freedom from Distractibility: a mix of memory and speed). This configuration of factor loadings was subsequently used in the confirmatory MG-CMSA, with the bold factor loadings estimated freely, and all other factor loadings fixed to zero.

3.3. Confirmatory factor analyses

3.3.1. Dutch sample

The results and fit statistics of the MG-CMSA on the Dutch data are presented in Table 4.

3.3.1.1. First-order factor models. In model F_1 we tested for configural invariance: a factor model with three correlated factors was fitted in boys and girls separately, with INF, SIM, AR, VOC and COMP loading on the Verbal factor, OA, BD, PC and PA loading on the Performance factor, and AR, BD, CO, MA and DS on the Memory factor. All these factor loadings, which were estimated separately in the two sexes, were significant in both boys and girls. The modification indices (MIs) showed however that the fit of this baseline model could be improved substantially by allowing the residuals of OA and BD to correlate (MI=20 in boys, and MI=18 in girls). Note that such minor modifications of the Wechsler-model are not uncommon, and this specific link between OA and BD has been established before (e.g., Arnau & Thompson, 2000; Dolan et al., 2006; Ward, Axelrod, & Ryan,

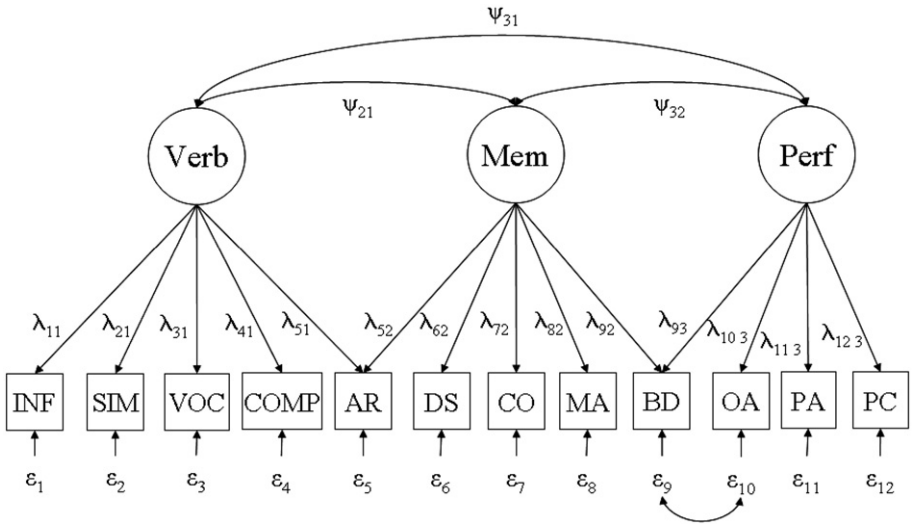

Fig. 1. First-order factor model for Dutch sample, where the λ's denote the regressions of the 12 subtests on the three factors, the Ψ's denote the correlations between the factors, and the ε's denote those parts of the variances in the subtests that are not predicted by the factors, i.e., the residual variances. VERB=Verbal factor, MEM=Memory factor, PERF=Performance intelligence, INF=Information, SIM=Similarities, AR=Arithmetic, VOC=Vocabulary, COMP=Comprehension, PC=Picture Completion, PA=Picture Arrangement, BD=Block Design, OA=Object Assembly, MA=Mazes, CO=Coding, DS=Digit span.

2000). Addition of these parameters to model F_{1a} in female and male samples resulted in a significant improvement of the fit ($\chi^2_{diff}(2)=26.84$, $p<.001$). Model F_{1a}, which is depicted in Fig. 1, will serve as the baseline model for all subsequent analyses. As the configuration of factor loadings and correlated residuals was identical in boys and girls, configural invariance across sex was established in the Dutch sample.

In model F_2 we tested for metric invariance by constraining the 14 factor loadings to be equal across sex, while the variances of the three common factors were freely estimated in the girls, and fixed to 1 in the boys for reasons of identification. These constraints did not result in a significant deterioration of the fit, compared to model F_{1a} ($\chi^2_{diff}(11)=9.43$, ns), so the factor loadings can be considered identical across sex.

Strong factorial invariance was tested in model F_3 by constraining the intercepts to be equal across sex, while the factorial means were fixed to zero in the boys, and estimated as latent mean differences in the girls. The fit of this model was however significantly worse than the fit of model F_2 ($\chi^2_{diff}(9)=65.66$, $p<.001$), implying that not all mean sex differences on the level of the subtests can be accounted for by differences on the level of the first-order factors. In view of the MIs, in model F_{3a}, it was decided to constrain all intercepts to be equal across sex, except the intercepts of INF, AR, and CO. Model F_{3a} did not fit significantly worse than model F_2 ($\chi^2_{diff}(6)=10.75$, ns). This means that strong factorial invariance was established for 9 of the 12 subtests. The sex differences on the subtests INF, AR and CO were too large to be accounted for by the first-order factors. So, in the sense discussed above, these three subtests may be viewed as biased within the common factor model. In all subsequent models, the subtest means of INF, AR, and CO were therefore estimated freely in each group, thereby effectively eliminating these subtests from the means model, while all other subtest means remained constrained to be equal across groups. Note that subtests that are biased with respect to their means can be retained in the model without consequence because, once these subtests' means have been relaxed (i.e., allowed to vary over sex), these indicators no longer contribute to the model for the means.

Strict factorial invariance was tested in model F_4 by constraining the residual variances plus the correlated residuals to be equal across sex. The fit of model F_4 was not significantly worse than the fit of model F_{3a} ($\chi^2_{diff}(13)=12.80$, ns), so the (correlated) residuals could be considered identical in boys and girls. The factor correlations and the factor means of this model are presented in Table 5. We find practically no sex difference with respect to Memory (−.01, s.e.), and small differences with respect to Verbal (−.26, s.e.) and Performance (−.19, s.e.). Fixing the first-order factor mean differences to be equal across sex did not result in a significant deterioration of model fit ($\chi^2_{diff}(3)=6.49$, $p=.09$), i.e., boys and girls did not differ significantly with respect to their means on the first-order factors. We note however, that the missingness present in the Dutch data may have reduced the statistical power to detect small factor mean differences between the sexes.

Table 5
Correlations between the first-order factors Verbal, Performance, and Memory for Dutch boys (below diagonal) and girls (above diagonal), and the means and standard deviations for boys and girls on the first-order factors

	Verbal	Performance	Memory
Verbal		.72	.49
Performance	.59	–	.60
Memory	.57	.39	–
Boys (N=350)			
Mean	0	0	0
SD	1	1	1
Girls (N=387)			
Mean	−.27	−.19	−.01
SD	1.04	1.01	.89
Effect size	−.26	−.19	−.01

Note. The means of the girls should be interpreted as deviations from the means of the boys, and were not significantly different from those of the boys (as tested in model S_3).

In sum, full measurement invariance was not tenable as the sex differences on INF, AR and CO were too large to be accounted for by the first-order factors. Partial measurement invariance was however tenable for the remaining 9 subtests, and all small (and non-significant) sex differences as observed on the level of these subtests could be described as (non-significant) differences on the level of the first-order factors. Although no significant mean differences were observed between boys and girls on the level of the first-order factors, a more parsimonious model for the means may identify a significant effect for sex. We therefore proceed in studying sex differences with respect to the second-order factor *g*.

3.3.1.2. Second-order factors models. In model S_1, '*g*', was introduced as a second-order factor for general intelligence. However, as there were only three first-order factors, this model with three first-order factors loading on 1 second-order factor was statistically equivalent to the model without a second-order factor, in which the first-order factors were simply correlated. The fit of model S_1 was thus identical to the fit of model F_4. Model S_1 is illustrated in Fig. 2.

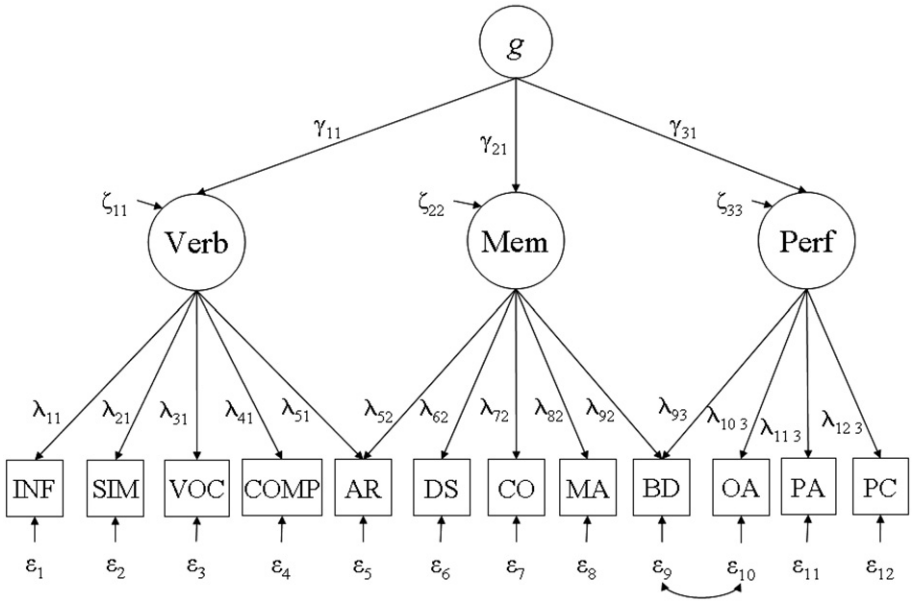

Fig. 2. The hierarchical factor model, where the λ's denote the regressions of the 12 subtests on the three first-order factors, the γ's denote the regressions of the first-order factor on the second-order factor *g*, and the ζ's and ε's denote those parts of the variances in the subtests and first-order factors that are not predicted by the first-order factors and the second-order factor, respectively. Note that the model is identical for the Dutch and Belgian sample, except that in the Belgian sample age-effects were regressed out on the level of the subtests (not drawn here for convenience). *g*=factor for general intelligence, VERB=Verbal factor, MEM=Memory factor, PERF=Performance intelligence, INF=Information, SIM=Similarities, AR=Arithmetic, VOC=Vocabulary, COMP=Comprehension, PC=Picture Completion, PA=Picture Arrangement, BD=Block Design, OA=Object Assembly, MA=Mazes, CO=Coding, DS=Digit span.

In model S_2, the second-order factor loadings were constrained to be equal across sex, and the variance of the second-order factor was fixed to 1 in the boys (for reasons of identification) and estimated freely in the girls. The fit of model S_2 was not significantly worse than the fit of model S_1 ($\chi^2_{diff}(2)=5.34$, ns), i.e., the factor loadings of the three first-order factors on *g* are identical across sex.

In model S_3, all first-order factor means were constrained to be zero in both boys and girls, while the mean of the second-order factor *g* was constrained to zero in boys for reasons of identification, and estimated freely in girls. The fit of the model did not deteriorate significantly as a result of these constraints ($\chi^2_{diff}(2)=.79$, ns), meaning that the sex differences with respect to the means of the first-order factors could be accounted for by the second-order factor.

Finally, model S_4, in which the second-order factor means were constrained to be identical for boys and girls (i.e., fixed to zero in both groups), did not fit the data significantly worse than model S_3 ($\chi^2_{diff}(1)=4.20$, $p=.04$). So we conclude that boys and girls do not

Table 6
Fit statistics Belgian sample

		CFI	RMSEA	χ^2_{diff}
F_1	Configural invariance	.97	.05	
F_{1a}	Configural invariance+residuals OA and BD correlated	.98	.04	F_{1a} vs. F_1: $\chi^2_{diff}(2)=48.01$, $p<.001$
F_2	Metric invariance	.99	.03	F_2 vs. F_{1a}: $\chi^2_{diff}(11)=4.89$, ns
F_3	Strong factorial invariance	.95	.06	F_3 vs. F_2:$\chi^2_{diff}(9)=112.90$, $p<.001$
F_{3a}	Strong factorial invariance, bar INF, AR and CO	.98	.04	F_{3a} vs. F_2: $\chi^2_{diff}(6)=17.05$, ns
F_4	Strict factorial invariance	.97	.04	F_4 vs. F_{3a}: $\chi^2_{diff}(13)=29.81$, $p<.01$
F_{4a}	Strict factorial invariance, bar INF	.98	.04	F_{4a} vs. F_{3a}: $\chi^2_{diff}(12)=23.31$, ns
S_1	Introduction 2nd order factor	.98	.04	S_1 is identical to F_{4a}
S_2	Metric invariance 2nd order factor	.97	.04	S_2 vs. S_1: $\chi^2_{diff}(2)=7.86$, ns
S_3	Strong factorial invariance 2nd order factor	.97	.04	S_3 vs. S_2: $\chi^2_{diff}(2)=4.21$, ns
S_4	Strict factorial invariance 2nd order factor	.97	.04	S_4 vs. S_3: $\chi^2_{diff}(1)<1$, ns

differ significantly with respect to *g*, i.e., with respect to general intelligence. We note again that the present missingness may have reduced the power to detect latent mean differences. For the record, we therefore note that the effect size for the difference in means between boys and girls on the second-order factor *g* was .25, which corresponds to a mean difference in favor of boys of 3.83 IQ points on a conventional IQ-scale.

3.3.2. Belgian sample

Because the age-range of the Belgian data was rather large (9.5–13 yrs) age was included in all models with age-effects regressed out on the level of the observed subtests. Note that all relations between age and the 12 subtests were estimated, i.e., the part of the model regarding the age-correction was saturated (fitted perfectly), and thus did not contribute to any misfit in the following models. The results and fit statistics of the MG-CMSA on the Belgian data are presented in Table 6.

3.3.2.1. First-order factor models. Configural invariance was tested in model F_1, i.e., a factor model with three correlated factors was fitted in boys and girls separately, with INF, SIM, AR, VOC and COMP loading on the Verbal factor, OA, BD, PC and PA loading on the Performance factor, and AR, BD, CO, MA and DS on the Memory factor. All these factor loadings proved significant in both groups. As in the Dutch sample, the fit of model F_1 could be improved by the introduction of an additional relation between the residuals of OA and BD (MI=21 in boys, MI=17 in girls). The fit of model F_{1a}, including these relations, was significantly better than the fit of model F_1 ($\chi^2_{diff}(2)=48.01$, $p<.001$). This baseline model F_{1a} is illustrated in Fig. 3. Note that with respect to the cognitive part of the model, model F_{1a} as fitted in the Belgian sample is identical to model F_{1a} as fitted in the Dutch sample.

Metric invariance was tested in model F_2, by constraining the 14 first-order factor loadings to be equal across sex, while the variances of the three first-order factors were estimated freely in the girls, and fixed to 1 in the boys for reasons of identification. As in the Dutch sample, the fit of model F_2 was not significantly worse than the fit of model F_{1a} ($\chi^2_{diff}(11)=4.89$, ns), suggesting that the loadings of the observed subtests on the primary factors of intelligence are identical in Belgian boys and girls.

Strong factorial invariance was tested in model F_3 by constraining the intercepts to be equal across sex, while

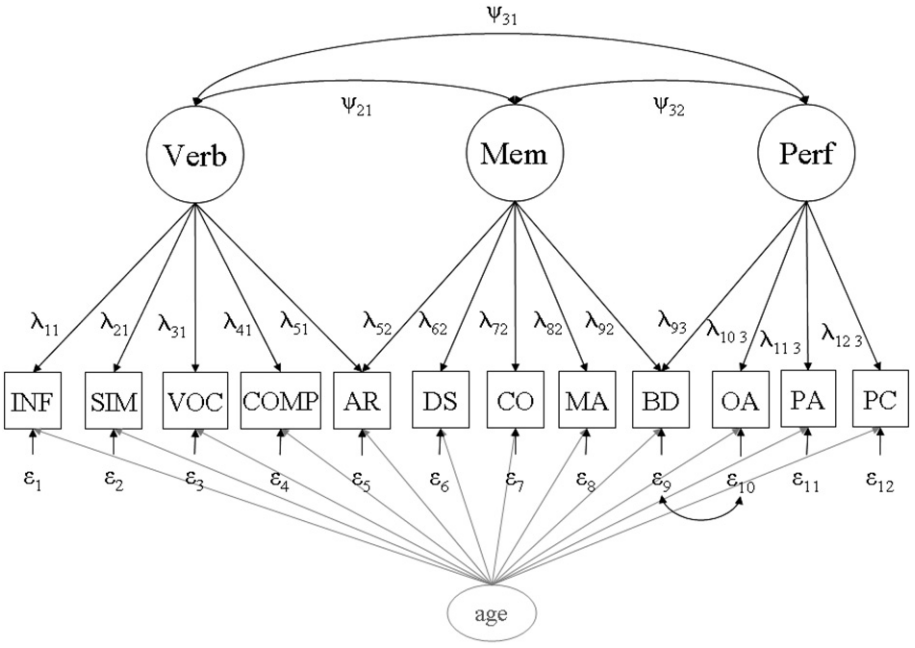

Fig. 3. First-order factor model for Belgian sample, where the λ's denote the regressions of the 12 subtests on the three factors, the Ψ's denote the correlations between the factors, and the ε's denote those parts of the variances in the subtests that are not predicted by the factors, i.e., the residual variances. Note that the Belgian model is identical to the Dutch model, except that in the Belgian sample age-effects were regressed out on the level of the subtests. VERB=Verbal factor, MEM=Memory factor, PERF=Performance intelligence, INF=Information, SIM=Similarities, AR=Arithmetic, VOC=Vocabulary, COMP=Comprehension, PC=Picture Completion, PA=Picture Arrangement, BD=Block Design, OA=Object Assembly, MA=Mazes, CO=Coding, DS=Digit span.

fixing the factorial means to zero in boys, and estimating them freely in girls. Like in the Dutch sample, these constraints led to a significant deterioration of the fit ($\chi^2_{diff}(9)=112.90$, $p<.001$), suggesting that not all sex differences on the level of the subtests can be accounted for by the first-order factors. As in the Dutch data, the modification indices suggested that INF, AR, and CO were causing this misfit. In model F_{3a}, all intercepts were constrained to be equal across sex, except the intercepts of INF, AR, and CO. The fit of model F_{3a} was just significantly worse than the fit of model F_2 ($\chi^2_{diff}(6)=17.05$, $p=.01$). The fit of this model would improve significantly ($\chi^2_{diff}(1)=9.53$, $p<.01$) if the mean of the subtest DS was freely estimated as well. However, free estimation of the means of subtests implies that these subtests are no longer part of the means model on the level of the first- or second-order factors. As the main aim of this study was to test whether the mean differences observed on the level of the subtests were indicative of mean differences on the level of the latent factors, we did not think it expedient to remove any more equality constraints on the subtest intercepts, as this would result in an undesirably stripped model. Since both the CFI and the RMSEA of model F_{3a} were good, and these results were comparable to those obtained in the Dutch sample, model F_{3a} was deemed reasonable. In all subsequent models, the subtest means of INF, AR and CO were therefore estimated freely in each group, thereby effectively eliminating these subtests from the model for the means. All other subtest means were constrained to be equal across groups.

In model F_4, strict factorial invariance was tested by constraining all (correlated) residual variances to be equal across sex. Unlike in the Dutch sample, the fit of model F_4 was significantly worse than the fit of model F_{3a} ($\chi^2_{diff}(13)=29.81$, $p<.01$). Modification indices showed that the misfit of model F_4 was mainly due to the residual of the subtest INF, which was not equal across sex. In model F_{4a}, all residual variances, bar the residual variance of INF, were constrained to be equal across gender. The resulting model did not fit significantly worse than model F_{3a} ($\chi^2_{diff}(12)=23.31$, $p.=025$). The inequality of the residual variance of INF implies that the reliability of this subtest is not equal in boys and girls. The factor correlations and the factor means of model F_{4a} are presented in Table 7. Note that fixing the first-order factor means to be equal across sex would not result in a significant deterioration of model fit ($\chi^2_{diff}(3)=4.52$, ns), i.e., boys and girls did not differ significantly with respect to their means on the first-order factors. As the Belgium samples are complete, and quite large, lack of power is unlikely to explain this absence of first-order factor mean differences.

Table 7
Correlations between the first-order factors Verbal, Performance, and Memory for Belgian boys (below diagonal) and girls (above diagonal), and the means and standard deviations for boys and girls on the first-order factors

	Verbal	Performance	Memory
Verbal	–	.70	.72
Performance	.71	–	.75
Memory	.62	.52	–
Boys (N=370)			
Mean	0	0	0
SD	1	1	1
Girls (N=391)			
Mean	−.09	−.04	−.46
SD	5.94	2.59	4.73
Effect size	−.02	−.02	−.13

Note. The means of the girls should be interpreted as deviations from the means of the boys, and were not significantly different from those of the boys (as tested in model S_3).

In sum, full measurement invariance was rejected as the sex differences on INF, AR, and CO could not be described by the first-order factors. Partial measurement invariance was however tenable for the remaining 9 subtests, for which the differences between boys and girls could, to reasonable approximation, be described as differences on the level of the first-order factors. Although sex differences were absent with respect to the first-order factor means, we proceeded in studying sex differences with respect to the second-order factor *g*. As mentioned before, absence of mean differences in a first-order factor model, does not guarantee the absence of mean differences in a more parsimonious model for the means.

3.3.2.2. Second-order factor models. A second-order factor for general intelligence, '*g*', was introduced in model S_1, but as there were only three first-order factors, model S_1 was equivalent to model F_4. The Belgian model S_1 is equivalent to the Dutch model S_1 (Fig. 2), except that in addition, age-effects were regressed out on the level of the subtests (see Fig. 3).

In model S_2, the second-order factor loadings were constrained to be equal across sex, while the variance of the second-order factor was freely estimated in the girls, and fixed to 1 in the boys for reasons of identification. Like in the Dutch data, the fit of model S_2 was not significantly different from the fit of model S_1 ($\chi^2_{diff}(2)=7.86$, $p=.02$), so the second-order factor loadings could be considered identical in boys and girls.

In model S_3, the means of the first-order factors were constrained to be zero in both boys and girls, while the mean of the second-order factor was fixed to zero in boys for reasons of identification, and estimated freely

in girls. Analogous to the Dutch results, the fit of the model did not deteriorate significantly, compared to model S_2 ($\chi^2_{diff}(2)=4.21$, ns), so the mean differences between boys and girls on the level of the first-order factors could be accounted for by the second-order factor.

Finally, the means of the second-order factor were constrained to be equal for boys and girls (i.e., fixed to zero in both groups) in model S_4. As the fit of model S_4 was, like in the Dutch data, not significantly worse than the fit of model S_3 ($\chi^2_{diff}(1)<1$, ns), we conclude that boys and girls do not differ with respect to *g*. This non-significance is reflected in the small effect size, which equaled .11, which corresponds to a mean difference in favor of boys of 1.58 IQ points on a conventional IQ-scale.

It is worth noting that additional confirmatory factor analyses on the Belgian data in which age was not regressed out, gave rise to almost identical results and thus to the same conclusions (tables of these results are available upon request).

4. Discussion

In this study, multi-group covariance and means structure analysis (MG-CMSA) was used to investigate sex differences on the WISC-R in Dutch and Belgian samples. In both samples, boys outperformed girls on the subtests Information (INF) and Arithmetic (AR), and girls outperformed boys on the subtest Coding (CO). The sex differences on these three subtests were too large to be accounted for by the first-order factors Verbal, Performance and Memory, i.e., were larger than was to be expected based on the first-order latent mean differences. In this sense, these three tests were 'biased' in the context of this measurement model, i.e., the sex differences on these subtests are genuine, but indicate differences in specific abilities rather than differences in the primary factors of intelligence that are distinguished in this model. INF, AR and CO were thus not measurement invariant with respect to sex, yet in both samples, measurement invariance with respect to sex could be established for all other subtests. That is, in both Dutch and Belgian boys and girls, scores on the remaining 9 subtests could be described as a function of scores on the latent factors, and the relations between the latent factors and the observed subtests were identical across sex.

When full measurement invariance is rejected and measurement invariance holds only partially (Byrne, Shavelson, & Muthén, 1989), one needs to decide whether it is meaningful to proceed with the group comparisons. As the means of the three biased subtests INF, AR and CO are effectively eliminated from the model, the meaning of the means of the first- and second-order factors will change. If one holds the opinion that the means of *g*, and the first-order factors, are not measured accurately without these three subtests in the model, further comparisons between boys and girls with respect to the first and second-order factors are useless. In that case, one should conclude that the present data cannot be used to establish sex differences in *g* as the test battery is not fully measurement invariant across sex. From a more pragmatic point of view, one could argue that, as partial measurement invariance is tenable, and all factors are still indicated by more than 2 measures, meaningful comparisons remain possible within the context of the revised model. In the present paper, we chose to continue with the group comparisons despite the partial measurement invariance.

At this point, we should note that we consider measurement invariance as an issue that concerns the relationships between observed variables (subtests) and latent variables, such as described in a measurement model. Thus, for instance, we view Information, Similarities, Vocabulary, Comprehension and Arithmetic as (at least hypothetical) indicators of the latent variable Verbal Comprehension. If any of these five subtests are not measurement invariant with respect to sex, this is problematic as it implies that boys and girls cannot be compared with respect to their scores on the latent factor Verbal Comprehension. Here one can resort to a partial measurement invariance model.

In contrast, we consider such strict psychometric criteria to be inappropriate when it comes to the relations of latent variables with other latent variables, such as the relations of first-order cognitive factors with the second-order factor '*g*'. First, the second-order factor model is not accorded the status of a measurement model but of a structural model, describing relations among latent variables or factors. Measurement invariance is not a criterion which latent factors are required to meet as they are usually not interpreted as "indicators of" or "measures of" other latent variables. Therefore, latent factors are usually not considered "biased" if their means cannot fully be accounted for by other latent factors. Second, the fact that the second-order model does not have the same psychometric status is evident for example in the evaluation of the first-order factor residuals. These first-order residuals are invariably accorded an important role in the hierarchical model, in contrast to the role of the residual terms of observed variables in the measurement model. For instance, *g* is hypothesized to be an important source of US black–

white differences in IQ test scores (Jensen, 1998). However, nobody subscribes to the idea that *g* is the only source of such differences, and mean differences in first-order factor residuals are invariably invoked to provide a full account of the observed mean differences between groups. For instance, Dolan (2000), having established measurement invariance, investigated a variety of latent variable models to account for observed mean differences in terms of latent mean differences. We do not interpret established mean differences in first-order factor residuals in terms of bias, because we do not view the structural model as a measurement model.

In sum, the study of measurement invariance is confined to the measurement model. Once measurement invariance of the observed subtest is fully or partially (but sufficiently) established, the interpretability of the latent factors is assured on the basis of the subtests which are measurement invariant. Subsequently, different structural models can be fitted, which describe the relations among the latent factors more or less parsimoniously.

In the present study, once the three biased subtests INF, AR and CO were effectively eliminated from the model for the means (by allowing the intercepts to differ over sex), sex differences on the level of the first-order factors Verbal, Performance and Memory were absent in both Dutch and Belgian data. Subsequently, sex differences on *g*, the second-order factor explaining all relations between the first-order factors, proved also statistically insignificant in both countries. Converted to the conventional IQ-scale (i.e., mean of 100, and standard deviation of 15), Dutch boys scored on average 3.83 IQ-point higher than Dutch girls, and Belgian boys scored on average 1.58 IQ points higher than Belgian girls. Note that the results for the Belgian sample were robust for the effect of age, i.e., inclusion of the effect of age in our models did not change any of the above conclusions with respect to sex differences.

The mean difference of 3.83 IQ-points between the Dutch boys and girls (with a *p*-value of .04) may be interpreted to indicate a trend towards male advantage with respect to general intelligence. Yet, given the absence of sex differences for the first-order factor means in both the Dutch and the Belgian sample, and given the clear absence of sex differences on the second-order factors in the Belgian sample, we prefer the conclusion that, in the age-range between 9 and 13, sex differences are present with respect to certain specific cognitive abilities (i.e., general knowledge, arithmetic and memory/speed), but absent with respect to the latent factors underlying intellectual performance. However, due to the missingness on 6 of 12 subtests for a subset of the Dutch sample, the power to detect sex differences in *g* for an observed effect size as small as .25 was not very high in the Dutch sample; about .60, given the present sample sizes and an α of .01 (see Dolan, van der Sluis & Grasman, 2005, for discussion of the influence of missingness on power in the context of structural equation modeling). Replication of the present results in a sample of children within the same age-range is therefore desirable.

This study illustrates the advantage of MG-CMSA. First, we were able to establish measurement invariance with respect to 9 out of 12 subtests, which ensures the comparability of factors and factor means across sex. Second, MG-CMSA allowed us to select those subtests for which sex differences were larger than was to be expected given the factor model. The sizable sex differences on the subtests INF and CO, and sometimes AR, are at present well documented (e.g., Born & Lynn, 1994; Jensen & Reynolds, 1983; Lynn, Fergusson, et al., 2005; Lynn & Mulhern, 1991; Lynn, Riane, et al., 2005). Our finding that, in both Dutch and Belgian data, these subtests were biased in the context of the assumed measurement model, suggests sex differences with respect to specific cognitive abilities. After elimination of these subtests from the means model, sex differences on the first- and second-order factors were small and statistically insignificant in both the Dutch and the Belgian sample. Finally, using MG-CMSA, one has the clear benefit of knowing that alternative hypotheses are tested within the context of a model that is known to provide an accurate description of the data.

One question of interest is whether the results with respect to sex differences would have turned out differently if we had chosen another measurement model. For example, what would the conclusion have been if we had used a bi-factor model (i.e., a model in which *g* is modeled as a first-order factor with loadings on all subtests, and Verbal, Memory and Performance are modeled as residual factors for specific abilities that are uncorrelated to *g*) rather then a hierarchical factor model? For the Dutch sample, the conclusions would have been exactly the same: measurement invariance holds for all subtests bar INF, AR and CO, and boys and girls do not differ with respect to any of the latent factors. For the Belgian sample, the assumption of equal factor loadings was untenable in the context of the bi-factor model ($\chi^2(22) = 50.00$, $p < .01$). If we, for the sake of argument, ignore the statistical significance of this difference (arguing for example that the actual difference between the factor loadings in boys and girls is small, and only statistically significant due to the large sample size), all other conclusions are similar to those

obtained in the context of the hierarchical factor model: INF, AR and CO are biased, and boys and girls do not differ significantly with respect to any of the factor means. The results obtained in the context of the hierarchical factor model thus seem robust. We preferred to test for sex differences in the context of the hierarchical factor model not only because it is more parsimonious than the bi-factor model, but also because it is consistent with the standard conception of *g* as a higher order factor. That is, we consider general intelligence or *g* to be the factor underlying all cognitive abilities, and do not adhere to the conceptualization of Verbal, Memory and Performance as specific cognitive abilities that are independent of general intelligence.

Another question of interest is whether twin-samples can be considered to be representative of the population. In general, twins are born in all strata of society, and are somewhat more willing than average to participate in research projects (Martin, Boomsma, & Machin, 1997). The issue whether the IQ of twins is on average somewhat lower than that of children born as singletons remains to be settled (e.g., Derom et al., 2005; Kallman, Feingold, & Bondy, 1951; Posthuma, de gues, Bleichrodt, & Boomsma, 2000; Ronalds, de Stravola, & Leon, 2005). However, there is as yet no reason to believe that the differences between the sexes should be different among boys and girls born as a twin. Furthermore, the present results concerning the sex differences on the subtests INF, CO and AR are consistent with previous studies (e.g., Born & Lynn, 1994; Jensen & Reynolds, 1983; Lynn, Fergusson, et al., 2005; Lynn & Mulhern, 1991; Lynn, Riane, et al., 2005). Also, the sex differences observed in the full Belgian and Dutch twin-samples correspond largely to the differences observed in the subsamples of opposite-sex twins, where familial confounders like SES are perfectly controlled for.

The difference between boys and girls with respect to *g*, while statistically insignificant, was somewhat smaller in the present study compared to previous studies, especially in the Belgian sample. This is probably due to elimination of the biased subtests from the model of the means; rather than concluding that sex differences are present on the level of latent factors, we conclude that boys and girls differ with respect to the specific cognitive abilities measured by the subtests INF, AR and CO. It is however also possible that our results differ from those reported in previous studies because of the more limited age-range of our samples. As most previous studies were performed on the full WISC-R standardization samples, with age ranging from 6 to 16 years, but without statistically correcting for possible age-effects, the reported sex-effects may not generalize to any sample of more limited age-range. As stated above, it even remains to be seen whether the factor structure of the WISC-R is stable between age 6 and 16, and whether the sex-effects reported for the (factor) means are stable across age. Lynn (1994, 1999) argued that researchers studying sex differences in cognitive ability in children, adolescents and young adults, have failed to take into account the fact that females on average mature somewhat earlier than males, thereby systematically underestimating the differences between the sexes. According to Lynn, the male advantage over females becomes visible once males and females reach adulthood, and maturational advantages of females cease to exist. Some support for this differential maturation hypothesis was recently offered by Colom and Lynn (2004), who, unfortunately, failed to test for measurement invariance across age and sex before comparing the scores of boys and girls. The present null-findings with respect to sex differences in *g* in children aged 9 to 13 years old do not contradict the expectations following from this developmental viewpoint. Yet, the fact that our results regarding sex differences were robust against age-effects, suggests that the present data showed little evidence for differential developmental trajectories for males and females. Whether samples with wider age-ranges show these hypothesized differential developmental trajectories in appropriate statistical analysis certainly merits further study.

In a meta-analysis of 57 studies on sex differences on the Raven Standard and Advanced Progressive Matrices in the general population, Lynn and Irwing (2004) reported the absence of sex differences before age 15, but a male advantage of about 5 IQ points from 15 years onwards. A meta-analysis performed on 22 studies in university students, confirmed the finding of this 5 IQ point difference in favor of males in young adults (Irwing & Lynn, 2005). Unfortunately, however, in all these studies, the Raven was analyzed as a unidimensional instrument, and item and subtests scores were simply added, and directly compared across groups. Recent studies on the dimensionality of the Raven have shown repeatedly the presence of more than one factor (e.g., Mackintosh & Bennett, 2005; Vigneau & Bors, 2005, Van der Ven & Ellis, 2000), and these different factors underlying performance on the Raven should be considered in studies on group/age differences. Although Lynn, Allik and Irwing (2004) and Mackintosh and Bennett (2005), did address the multi-dimensionality of the Raven, they did not test for measurement invariance in their respective studies on sex differences.

Two studies in adult samples, in which MG-CMSA was employed to investigate sex differences on the WAIS, showed that, after elimination of the biased subtests, males outperformed females on the WAIS first-order factors Working Memory and Perceptual Organization, while females outperformed males on Perceptual Speed, and no sex differences were found for the factor Verbal Comprehension (Dolan et al., 2006; Van der Sluis et al., 2006). According to these authors, the observation of positive as well as negative sex differences on four positively correlated factors does not fit well with the idea of one factor (i.e., the second-order factor *g*) being the only source or cause of the observed sex differences. So although these data did not provide evidence for sex differences in *g*, in adults sex differences were clearly present on the level of the first-order factors. The finding that sex differences on the first-order factors are absent in children (present study), while such differences have been observed in adults (Dolan et al., 2006, Van der Sluis et al., 2006), is in line with the idea proposed by Lynn (1994, 1999) that sex differences become more apparent with age.

In sum, although the Raven on the one hand, and the WISC-R and the WAIS on the other are not directly comparable with regard to number and nature of subtests, or number and nature of factors assumed to underlie performance on the subtests, the difference between the present results (i.e., no sex differences on the level of the latent factors) and the results in adult samples (i.e., sex differences are present on the level of the latent factors) could be indicative of an effect of age on the factor structure, or of an effect of age on the nature of sex differences in cognitive ability, such as hypothesized by Lynn (1994, 1999). Re-analysis of the WISC-R standardization data as analyzed previously (e.g., Born & Lynn, 1994; Jensen & Reynolds, 1983; Lynn, Fergusson et al., 2005; Lynn & Mulhern, 1991; Lynn, Riane, et al., 2005) using MG-CMSA, while accounting for age-effects, would therefore be a fruitful addition to the literature on sex differences in intelligence. As yet, the extent to which findings with respect to group differences in *g* on one measure of intelligence are generalizable to other intelligence tests is unknown. For example, correlations between the composite scores of the Raven and the WAIS range between .40 and .75 (e.g., Mackintosh, 1998). Although clearly substantial, these correlations cannot be taken to mean that Full Scale IQ scores on the WAIS and scores on the Raven represent identical constructs. Simultaneous analysis of *g* in the Raven and *g* in the WISC or the WAIS, using the appropriate statistical techniques (i.e., techniques that consider the underlying factor structure and the model's fit, and that test for measurement invariance) seems the only sensible way to study the comparability of *g*, and the comparability of mean differences in this *g*, in disparate intelligence test batteries. For example, see the study by Johnson, Bouchard, Krueger, McGue and Gottesman (2004), who report very high correlations between the second-order factors *g* calculated for the WAIS, the CAB, and a Hawaiian Battery, which includes the Raven. These results were obtained in a sample of adult males and females. However, sex differentiation was not addressed.

The present study is the first to examine measurement invariance and sex differences on the WISC-R using multi-group covariance and mean structure analysis. This resulted in a refined description of the nature of the sex differences in cognition in children aged 9 to 13. It was shown that measurement invariance was only partially tenable: in both the Dutch and the Belgian sample, the differences between boys and girls with respect to the specific abilities measured by the subtests INF, AR and CO were too large to be accounted for by the underlying latent factors. Once these biased subtests were removed from the model by estimating the intercepts free over sex, no differences were found with respect to the primary and secondary common factor means. The general conclusion is therefore twofold: 1) the WISC-R is only partially measurement invariant across sex in children aged 9–13; and 2) the main cognitive differences between boys and girls appear to concern specific abilities, and these sizeable differences do not seem to be attributable to differences in the second-order factor *g*. However, full measurement invariance across sex could not be established for the WISC-R. This psychometric deficiency of the WISC-R may detract from the power to detect the exact role of g, if any, in sex differences. However, the finding that Information, Coding and Arithmetic are biased is in itself an important empirical finding, which requires further study. Such studies could include the investigation of measurement invariance at the level of the items which comprise these subtests, using a suitable IRT model. In addition, as pointed out by a reviewer, techniques like functional imaging may prove useful in this context.

Still, on the basis of the present results, we conclude that explanations of sex differences in scores on IQ tests in the present age-range of 9 to 13 should be sought in specific abilities rather than in *g*.

Acknowledgements

The Belgian research was supported by the Marguerite-Marie Delacroix Foundation. The Dutch research

was supported by The Netherlands Organization for Scientific Research grants 575-25-012, 904-57-94 and 575-25-006. Preparation of this manuscript was financially supported by NWO/MaGW VIDI-016-065-318.

We heartily thank the reviewers for their constructive criticism.

References

Anderson, T., & Dixon, W. E. (1995). Confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Revised with normal and psychiatric adolescents. *Journal of Research on Adolescence*, *5*(3), 319−332.

Arnau, R. C., & Thompson, B. (2000). Second-order confirmatory factor analysis of the WAIS-III. *Assessment*, *7*(3), 237−246.

Bartels, M., Rietveld, M. J. H., van Baal, G. C. M., & Boomsma, D. I. (2002). Genetic and environmental influences on the development of intelligence. *Behavior Genetics*, *32*(4), 237−249.

Bartels, M., van Beijsterveldt, C. E. M., Stroet, T. M., Hudziak, J. J., Boomsma, D. I., & Young-Netherlands Twin Register (Y-NTR) (2007). A longitudinal multiple informant study of problem behavior. *Special Issue Twin Research and Human Genetics*, *10* (1)(Feb).

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238−246.

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

Bollen, K., & Long, J. S. (Eds.) (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

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

Boomsma, D. I., Vink, J. M., van Beijsterveldt, T. C. E. M., de Geus, E. J. C., Beem, A. L., Mulder, E. J. C. M., et al. (2002). Netherlands Twin Register: A focus on longitudinal research. *Twin Research*, *5*(5), 401−406.

Born, M. P., & Lynn, R. (1994). Sex differences on the Dutch WISC-R: A comparison with the USA and Scotland. *Educational Psychology*, *14*(2), 249−254.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136−162). Newbury Park, CA: Sage.

Burton, D. B., Sepehri, A., Hecht, F., VandenBroek, A., Ryan, J. J., & Drabman, R. (2001). A confirmatory factor analysis of the WISC-iii in a clinical sample with cross-validation in the standardization sample. *Child Neuropsychology*, *7*(2), 104−116.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456−466.

Centraal Bureau voor de Statistiek (2002). *Enquête beroepsbevolking 1999. Standaard beroepsclassificatie 1992*. . Heerlen: Centraal Bureau voor de Statistiek.

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.

Dai, X. Y., & Lynn, R. (1994). Gender differences in intelligence among Chinese children. *Journal of Social Psychology*, *134*(1), 123−125.

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

Derom, C., Thiery, E., Derom, R., van Gestel, S., Jacobs, N., Vlietinck, R., et al. (2005). The cognitive costs of being a twin girl. *Rapid Responses, British Medical Journal*. http://bmj.bmjjournals.com/cgi/eletters/bmj.38633.594387.3Av1

Derom, C., Vlietinck, R., Thiery, E., Leroy, F., Fryns, J. P., & Derom, R. (2002). The East Flanders Prospective Twin Survey (EFPTS). *Twin Research*, *5*, 337−341.

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. (2006). Multi-group covariance and mean structure modeling of the relationship between WAIS-III common factors and gender and educational attainment in Spain. *Intelligence*, *34*(2), 193−210.

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.

Dolan, C. V., van der Sluis, S., & Grasman, R. (2005). A note on normal theory power calculation in structural equation modeling with data missing completely at random. *Structural Equation Modeling*, *12*(2), 245−262.

Donders, J. (1993). Factor structure of the WISC-R in children with traumatic brain injury. *Journal of Clinical Psychology*, *49*(2), 255−260.

Grégoire, J. (2000). *L'évaluation clinique de l'intelligence de l'enfant. Théorie et pratique du WISC-III*. . Bruxelles: Mardaga.

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.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1−55.

Huberty, T. (1987). Factor analysis of the WISC-R and the adaptive behavior scale-school edition for a referral sample. *Journal of School Psychology*, *25*, 405−410.

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.

Jacobs, N., Van Gestel, S., Derom, C., Thiery, E., Derom, R., Vernon, P., et al. (2001). Heritability estimates of intelligence in twins: Effect of chorion type. *Behavior Genetics*, *31*, 209−217.

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.

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one *g*: Consistent results from three test batteries. *Intelligence*, *32*, 95−107.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294−316). Newbury Park, CA: Sage.

Kallman, F. J., Feingold, L., & Bondy, E. (1951). Comparative adaptational, social, and psychometric data on the life histories of

senescent twin pairs. *American Journal of Human Genetics*, *3*, 65−73.

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. New York: the Guilford Press.

Kush, J. C., Watkins, M. W., Ward, T. J., Ward, S. B., Canivez, G. L., & Worrell, F. C. (2001). Construct validity of the WISC-III for white and black students from the WISC-III standardization sample and for black students referred for psychological evaluation. *School Psychology Review*, *30*(1), 70−88.

Lawley, D., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. London: Butterworth.

Little, T. D. (1997). Mean and covariance structures (MACS) analysis of cross-cultural data: practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53−76.

Loos, R., Derom, C., Vlietinck, R., & Derom, R. (1998). The East Flanders Prospective Twin Survey (Belgium): A population-based register. *Twin Research*, *1*(1), 167−175.

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

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence*, *32*, 411−424.

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.

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

Lynn, R., & Mulhern, G. (1991). A comparison of sex difference on the Scottish and American standardisation samples of the WISC-R. *Personality and Individual Differences*, *12*(11), 1179−1182.

Lynn, R., Riane, A., Venables, P. H., Mednick, S. A., & Irwing, P. (2005). Sex differences on the WISC-R in Mauritius. *Intelligence*, *33*, 527−533.

Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford: Oxford University Press.

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's matrices measure? An analysis in terms of sex differences. *Intelligence*, *33*, 663−674.

Martin, N., Boomsma, D., & Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics*, *17*(4), 387−392.

Meesters, C., van Gastel, N., Ghys, A., & Merckelbach, H. (1998). Factor analysis of the WISC-R and the K-ABC in a Dutch sample of children referred for learning disabilities. *Journal of Clinical Psychology*, *54*(8), 1053−1061.

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.

Muthén, L. K., & Muthén, B. O. (2005). *Mplus user's guide* (3rd ed.). Los Angeles, CA: Muthén & Muthén.

Oh, H. J., Glutting, J., Watkins, M. W., Youngstrom, E. A., & McDermott, P. A. (2004). Correct interpretation of latent versus observed abilities: Implications from structural equation modeling applied to the WISC-R and the wait linking sample. *Journal of Special Education*, *38*(3), 159−173.

Polderman, T. J. C., Stins, J. F., Posthuma, D., Gosso, M. F., Verhulst, F. C., & Boomsma, D. I. (2006). The phenotypic and genotypic relation between working memory speed and capacity. *Intelligence*, *34*(6), 549−560.

Posthuma, D., de Geus, E. J. C., Baaré, W. F. C., Hulshoff Pol, H. E., Kahn, R. S., & Boomsma, D. I. (2002). The association between brain volume and intelligence is of genetic origin. *Nature Neuroscience*, *5*(2), 83−84.

Posthuma, D., de geus, E. J. C., Bleichrodt, N., & Boomsma, D. I. (2000). Twin-singleton differences in Intelligence? *Twin Research*, *3*, 83−87.

Rebollo, I., de Moor, M. H. M., Dolan, C. V., & Boomsma, D. I. (2006). Phenotypic factor analysis of family data: Correction of the bias due to dependency. *Twin Research and Human Genetics*, *9*(3), 367−376.

Rietveld, M. J. H., van der Valk, J. C., Bongers, I. L., Stroet, T. M., Slagboom, P. E., & Boomsma, D. I. (2000). Zygosity diagnosis in young twins by parental report. *Twin Research*, *3*(3), 134−141.

Ronalds, G. A., de Stravola, B. L., & Leon, D. A. (2005). The cognitive costs of being a twin: evidence from comparisons within families in the Aberdeen children of the 1950s cohort study. *British Medical Journal*, *331*, 1306.

Rushton, J. P., & Jensen, A. R. (2003). African-white IQ differences from Zimbabwe on the Wechsler Intelligence Scale for Children-Revised are mainly on the *g*-factor. *Personality and Individual Differences*, *34*, 177−183.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp. 233−247). London: Kluwer Academic Publishers.

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

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23−74.

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.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*, 31−65.

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

Van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, *29*, 45−64.

Van Haasen, P. P., De Bruyn, E. E. J., Pijl, Y. J., Poortinga, Y. H., Lutje-Spelberg, H. C., Vander Steene, G., et al. (1986). *Wechsler Intelligence Scale for Children-Revised, Dutch Version*. Lisse, The Netherlands: Swets & Zetlinger B.V.

Vigneau, F., & Bors, D. A. (2005). Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement*, *65*(1), 109−123.

Ward, L. C., Axelrod, B. N., & Ryan, J. J. (2000). Observations on the factor structure of the WAIS-R. *Assessment*, *7*(1), 79−86.

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.

Widaman, K. F., & Reise, S. P. (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.

Wright, D., & Dappen, L. (1982). Factor analysis of the WISC-R and the WRAT with a referral population. *Journal of School Psychology*, *20*(4), 306−312.