Intelligence 38 (2010) 101–110

Contents lists available at ScienceDirect

Intelligence

Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied?

Ricarda Steinmayr [a,*], André Beauducel [b], Birgit Spinath [a]

[a] University of Heidelberg, Department of Psychology, Hauptstrasse 47-51, D-69117 Heidelberg, Germany
[b] University of the Federal Armed Forces, Personality Psychology, Holstenhofweg 85, D-22043 Hamburg, Germany

ARTICLE INFO

Article history:
Received 6 May 2009
Received in revised form 11 August 2009
Accepted 17 August 2009
Available online 9 September 2009

Keywords:
Sex differences
Latent cognitive variables
Factor scores

ABSTRACT

Recently, different methodological approaches have been discussed as an explanation for inconsistencies in studies investigating sex differences in different intelligences. The present study investigates sex differences in manifest sum scores, factor score estimates, and latent verbal, numerical, figural intelligence, as well as fluid and crystallized intelligence as measured by the German Intelligence-Structure-Test 2000-R (IST 2000-R; Liepmann, Beauducel, Brocke, & Amthauer, 2007). The not population-representative sample consisted of 977 German 11th and 12th graders enrolled in a "Gymnasium" (551 female; mean age: $M = 16.70$; $SD = 0.65$) who completed the IST 2000-R. Sex differences in fluid and crystallized intelligence were not influenced by the method applied with men performing better than women. However, extent and direction of sex differences in verbal, numerical, and figural intelligence differed by the method applied. Whereas there was a male advantage in all three factors measured as manifest sum scores, women performed better in verbal intelligence as measured by factor scores or as latent variables. Effect sizes of sex differences in numerical and figural intelligence were also greatly reduced when applying the latter two methods. Results are discussed with regard to their theoretical and practical implications.

1. Introduction

Sex differences in intelligence have been extensively investigated ever since the first intelligence tests have been introduced. Despite the substantial attention the topic has received, it is far from being thoroughly illuminated as research on sex differences in intelligence has partly produced inconsistent results concerning the presence, magnitude, and direction of the effects. Possible explanations for these inconsistencies are developmental effects (cf. Lynn, 1999), selective samples (cf. Dykiert, Gale, & Deary, 2009), and the measures used (cf. Lynn, 1999). A further explanation for the inconsistencies might be the methodological approach applied to the investigation of group differences (cf., e.g., Keith, Reynolds, Patel, & Ridley, 2008). Recently sex differences in intelligence have been investigated by multivariate latent variable approaches rather than comparing manifest intelligence test scores. An advantage of a latent variables approach is that it allows conclusions about sex differences in underlying, pure intelligence factors. Contrary to this, comparing manifest intelligence test scores might yield misleading results about the true nature of sex differences. Nevertheless, it is important to investigate sex differences in measured manifest intelligence because real world decisions, such as selection for jobs, are based on manifest intelligence test scores. The present study examines sex differences in manifest sum scores, factor score estimates, and latent verbal, numerical, figural intelligence, as well as fluid and crystallized intelligence as measured by the German Intelligence-Structure-Test 2000-R (Liepmann, Beauducel, Brocke, & Amthauer, 2007). Thus, the present study aims to compare the three methodological approaches and their impact on the emergence of sex differences.

* Corresponding author.
E-mail addresses: ricarda.steinmayr@psychologie.uni-heidelberg.de (R. Steinmayr), beauduce@hsu-hh.de (A. Beauducel), birgit.spinath@psychologie.uni-heidelberg.de (B. Spinath).

0160-2896/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2009.08.001

1.1. Sex differences in cognitive abilities

Most attention in research on sex differences in intelligence has been paid to sex differences in general intelligence. In accordance with two pioneers of intelligence testing, Terman (1916) and Wechsler (1944), it has long been proposed that there are no sex differences in general intelligence (e.g., Brody, 1992). However, this view has been challenged by different authors (e.g., Lynn, 1999; Nyborg, 2003), claiming that there are differences in general intelligence favouring adolescent and adult males. There is support for both views. Some authors found that men exceed women in general intelligence (e.g., Irwing & Lynn, 2005; Jackson & Rushton, 2006; Lynn & Irwing, 2004a,b, 2008; Nyborg, 2005) whereas others reported no sex differences (e.g., Aluja-Fabregat, Colom, Abad, & Juan-Espinosa, 2000; Colom, García, Juan-Espinosa, & Abad, 2002; Deary, Thorpe, Wilson, Starr, & Whalley, 2003; Johnson & Bouchard, 2007; van der Sluis et al., 2008). Some studies even depict a female advantage in general intelligence (e.g., Keith et al., 2008; Reynolds, Keith, Ridley, & Patel, 2008).

Considering sex differences in broad cognitive abilities such as verbal, numerical, or figural intelligence, the findings by Terman (1916) and Wechsler (1944) still seem to hold for the present day. These authors found a male advantage in more complex numerical tasks as well as in tasks requiring figural abilities. On the other hand, women received higher scores in some verbal tasks and measures of perceptual speed. These effects can largely be found in the current overview of different meta-analyses on sex differences in cognitive abilities provided by Hyde (2005). The findings of sex differences in two further broad cognitive abilities, namely fluid and crystallized intelligence, are less consistent.

Horn (1988, p. 660) depicts fluid intelligence as a “fallible indicator of reasoning of several kinds, abstracting, and problem solving, when these qualities are acquired outside the acculturation process, through personal experience, and through learning that is not selectively restricted.” The author defines crystallized intelligence as follows: “The measured factor is a fallible indicator of the extent to which an individual has incorporated, through the systematic influence of acculturation, the knowledge and sophistication that can be referred to as the intelligence of a culture.” (Horn, 1988; pp. 658–659) Some authors found sex differences favouring men in fluid intelligence (Lynn & Irwing, 2002, 2004a,b), whereas others found none or inconsistent results with some tests favouring men and some favouring women (Colom & García-López, 2002; Keith et al., 2008; Reynolds et al., 2008). The picture for crystallized intelligence seems somewhat more homogeneous. Most studies demonstrate men's superior performance (e.g., Ackerman, Bowen, Beier, & Kanfer, 2001; Lynn, Irwing, & Cammock, 2002; Reynolds et al., 2008; van der Sluis et al., 2006) whereas others found no sex differences (e.g., Kaufman, Chen, & Kaufman, 1995).

Possible explanations for the inconsistent results in both general intelligence and broad intelligences have been offered. Whereas developmental changes (Lynn, 1999, 1994), recruitment strategies (Dykiert et al., 2009), and operationalization of general and broad intelligences by different tests (Colom & García-López, 2002; Lynn, 1999) have already been investigated and provide some explanation for the inconsistent results, newer explanations refer to methodological considerations when investigating sex differences.

1.2. Methodological considerations

Different methodological approaches have been mentioned as a further explanation for both inconsistent results in studies investigating sex differences in general intelligence and broad cognitive abilities (e.g., Rosén, 1995; van der Sluis et al., 2006, 2008). Methodological approaches to study sex differences vary greatly. Whereas meta-analyses tend to concentrate on studies investigating sex differences by means of the standardized sum scores (e.g., Lynn & Irwing, 2004a, 2008), other studies either investigate sex differences in cognitive abilities via factor score estimates, components from factor analysis or principal components analysis (for an overview, cf. Nyborg, 2003). More recently, sex differences in latent variables are investigated by means of structural equation modelling (e.g., Dolan et al., 2006; Keith et al., 2008; Maitland, Intrieri, Schaie, & Willis, 2000; Reynolds et al., 2008; Rosén, 1995; van der Sluis et al., 2006, 2008). Depending on the method, different results might emerge.

This shall first be illustrated by means of general intelligence. According to Spearman (1904), each single test score in a multifactorial intelligence test represents general intelligence plus specific abilities and skills and measurement error specific to the particular test. From this it follows, that general intelligence (*g*) is based on the correlations among test scores and thus represents their shared variance. The standardized composite score (IQ) rests on a summation of the single test scores and thus represents *g* plus the various specific abilities and skills plus measurement error. Moreover, the summation of test scores does not take into account the relevance of the different tasks for general intelligence. Thus, depending on the method applied different results might emerge. This is in line with the findings by Colom et al. (2002) who found no sex differences in general intelligence defined as Spearman's *g* whereas males had an advantage in general intelligence defined as the sum of cognitive abilities. Consequently, when comparing any groups on an IQ score as a proxy for *g* the analysis does not allow substantive conclusions as to whether these groups differ in *g*. Mean group differences may also be due to group differences on the subtest level, indicate sex differences in broader cognitive abilities such as verbal, numerical or figural intelligence or may just reflect specifics in the measurement error related to the group but unrelated to intelligence.

The same considerations pertain to broad cognitive abilities such as verbal, numerical, or figural intelligence as well as fluid or crystallized intelligence (gf and gc). Group comparisons of these abilities are also often made on the basis of subtest scores or scales based on different subtests representing the same broad cognitive ability. According to Spearman (1904), broad cognitive ability test scores represent the focal particular broad cognitive ability, *g*, and measurement error. Johnson and Bouchard (2007) demonstrated that sex differences on subtest level greatly increased when partialling out general intelligence. The authors concluded that sex differences in specific cognitive abilities are overshadowed by *g*. Consequently it is impossible to speak about group differences in any of these abilities without taking into consideration the other abilities. For example, as reported above it is generally found that women exceed men in verbal abilities (e.g., Feingold, 1988; Halpern & LaMay, 2000). However, Hyde and Linn (1988) performed a meta-analysis of 165 studies of sex differences in verbal ability and concluded that the differences in overall verbal abilities are too small to be of practical relevance. These authors found larger effects for more narrow verbal tasks. Females obtained higher scores in tests of general verbal ability, anagrams, and speech production, but boys performed better on tasks assessing verbal analogies. Verbal analogies are often used as a measure of verbal reasoning and might be loaded by general reasoning. Adolescent boys or men have most often been found to exceed girls or women in abstract reasoning (e.g., Feingold, 1988; Lynn, 1992; Lynn, Allik, Pullmanm, & Laidra, 2004, found no sex differences in abstract reasoning). Thus, sex differences in tests measuring verbal reasoning might only be due to boys' or men's advantage in abstract reasoning which might overshadow girls' or women's real advantage in verbal materials. As the verbal reasoning tasks were also included in the examination of sex differences in overall verbal ability, the masking effect might also explain why Hyde and Linn (1988) did not find larger effects.

From this argument it follows that manifest test scores do not optimally represent different cognitive abilities because they are loaded with other, non-focal cognitive abilities or even measurement error. Thus, it is necessary to apply other methods such as factor score estimates or latent covariance analysis to draw conclusions on which underlying constructs men and women really differ. These methodological approaches may allow for more accurate conclusions about sex differences in cognitive abilities as they allow for differentiation of various variance sources of observed scores. Structural equation modelling with latent variables most clearly avoids the problem of confounding constructs with measurement variance and specific variance. Therefore, sex and age differences found in latent variables as they are conceived in structural equation models most directly avoid incorrect inferences based on measurement error and specific variance. On the other hand, differences found in latent variable models cannot be generalized to any type of individual scores, because the latent variable models do not contain scores on an individual basis. The most precise method for generating individual scores from latent variables is the computation of factor score estimates. Because manifest test scores are relevant whenever intelligence tests are used in applied contexts, it seems important to investigate sex and age differences not only within the latent variable models but also at the level of factor score estimates, which are more precise indicators of the latent variables than the observed manifest scores and unweighted scale scores. The group means of factor score estimates and the group means computed directly in latent variable models are not necessarily identical, because of the indeterminacy of factor score estimates. The indeterminacy problem implies that different factor score estimates can be computed for one and the same latent variable model (Lawley & Maxwell, 1971). Therefore, the choice of factor score estimates would have an impact on the group means. In the present case, it was regarded as especially important that the factor score estimates have the same intercorrelations as the latent variables themselves. Therefore, McDonald's (1981) factor score estimate was used as a basis for the computation of group means.

We are not aware of a study that has compared the three methodological approaches (manifest sum scores, latent variables, factor score estimates) with regard to the emergence of sex differences within one sample. This procedure is, however, necessary in order to draw conclusions as to the impact of the three methods on the magnitude and direction of sex differences as, otherwise, the possibility cannot be excluded that found inconsistencies in sex differences are due to samples differing in age or selection and/or to different measures.

1.3. Aim of the present study

The aim of the present study was to investigate sex differences in intelligence test scores obtained by three different methods. We compared the association between sex and cognitive abilities for scales based on manifest sum scores of the given intelligence test, for corresponding latent variables in a structural equation model, and for factor score estimates corresponding to the latent variables. We chose these methods out of all available possibilities as they are commonly used in studies investigating sex differences. Note that the presented results are limited to these methods and to the use of the specific intelligence test.

A large sample of high school students was investigated. We focused on a school-based sample because the chances of recruiting representative samples are higher in schools than in other settings (Dykiert et al., 2009). Testing in schools is often mandatory or the "costs" to participate are not as high as in other settings because testing takes place during daily routine. Because age has been identified as an important variable to be considered in research on sex differences (cf. Hyde, 2005), we investigated a small age range for which Liepmann et al. (2007) did not find age to influence test performance when applying the same test. However, some limitations to the sample must be mentioned as they are important for interpreting the results. The sample was not population representative but represents the typical population of one specific school type. Gymnasium is the most demanding school in the German school system. Consequently, the sample was recruited from the higher end of the ability distribution and was most comparable to a college student population. Pupils are selected for Gymnasium according to their grades in elementary school. Because girls get better grades in school, more female than male students attend the Gymnasium. As a consequence, males were found to perform better on intelligence tests than females (Steinmayr & Spinath, 2008). Furthermore, it might well be that the development of sex differences is not fully completed at the investigated age range. Thus, the following results do not allow for conclusions about sex differences in the general population. As stated above, the primary interest of the paper was to investigate whether the methods used to estimated sex differences affect the results obtained within a given sample.

We used structural equation modelling to examine the association between sex and different broad cognitive abilities. The measurement and theoretical models were based on a well validated intelligence model (Beauducel, Brocke, & Liepmann, 2001; Schulze, Beauducel, & Brocke, 2005) that also served as the basis for the development of the intelligence test used here. The model contained facets for fluid and crystallized intelligence as well as facets for verbal, numerical, and figural intelligence. The latter three factors represented content intelligence factors. In this model, fluid and crystallized intelligence were conceived independently from the verbal, numerical or figural content (Beauducel, Liepmann, Felfe, & Nettelnstroth, 2007). As this model avoided overshadowing of the content factors with fluid and crystallized intelligence, this model might give new insights into sex differences of the focal intelligences. Due to the reported inconsistencies concerning sex differences in the focal cognitive abilities and the specifics of our sample, we abstained from formulating concrete hypotheses. The following research questions were investigated:

1) Are there sex differences in the five investigated broad cognitive variables (verbal, numerical, figural, crystallized and fluid intelligence) in manifest sum scores, factor score estimates, and latent variables in the present sample?
2) Do the effect sizes of sex differences in manifest sum scores, factor score estimates, and latent broad cognitive variables differ?

2. Method

2.1. Participants

We investigated 977 students (551 female, 426 male) recruited from different schools in two federal states in Germany. All students were enrolled in a Gymnasium. Graduating from this type of school is one possible route to receive the mandatory school leaving certificate required for university enrolment. Typically, the female:male ratio is about 55:45 (Statistisches Bundesamt, 2008). Thus, the share of males and females of the present sample was representative for this student population. All students either attended 11th or 12th grade. Age ranged from 16 to 18 years with a mean age of $M = 16.70$ ($SD = .65$).

2.2. Measures

The I-S-T 2000 R included nine reasoning tasks and a knowledge test. The reasoning tasks were subdivided into three verbal, three numerical, and three figural reasoning tasks (Liepmann et al., 2007). Examples of the tasks can be found in Schulze et al. (2005). Each verbal, numerical, and figural reasoning task consisted of 20 items. In the first verbal task, “Sentence Completion” (SC), participants had to choose one out of five words in order to complete a sentence correctly. In the second verbal task, “Verbal Analogies” (VA), participants had to choose one out of five words in order to complete a verbal analogy correctly. In the “Verbal Similarities” (VS) task groups of six words were presented to the participants who were requested to choose the two words included that depicted the most similar content of all words. The first numerical task, “Calculations” (CA) consisted of simple arithmetic operations. In the second numerical task, “Number Series” (NS), participants had to write down the next number corresponding to a rule, which had to be identified. In the “Signs” task (SI), participants had to insert arithmetical signs into equations in order to complete them correctly. “Abstract Pieces” (AP) was the first figural reasoning task. It required that participants choose one out of five geometric figures, which can be composed by smaller pieces. The “Cubes” (CU) required the participants to choose one out of five cubes, which represented a rotated target cue. The third figural reasoning task, “Matrices” (MA), corresponded to abstract figural matrices as they are often used to assess fluid intelligence. The knowledge test was comprised of 84 questions covering the domains of geography/history, business, science, mathematics, arts, and daily life. The questions were formulated in verbal, numerical, and figural contents. Each of the three contents comprised 28 knowledge items. To test the factorial structure of the test, the test manual suggested forming 3 verbal, 3 numerical, and 3 figural knowledge aggregates comprising 9 or 10 specific items. In the present study manifest verbal, numerical, and figural intelligence scores were computed as sum scores of the corresponding reasoning tasks and knowledge aggregates. A knowledge summation score representing crystallized intelligence at a manifest level was computed by summing up all knowledge items. A sum score of the nine reasoning subtests was computed in order to represent fluid intelligence at a manifest level. All subtests were summed up to represent a measure of overall cognitive abilities.

2.3. Procedure

Testing took place during a regular school day and took about 3 h. Students could either decide to take part in the testing or to work on an extra-assignment during the testing time in the teachers' room. All students that came to school at the testing days participated and none chose to work on the extra-assignment. Testing was conducted in groups of about 20 students and tests were given by trained students and research assistants. First, we administered the short form of the basic module of the IST 2000-R which lasted about 95 min. Second, after a break of 15 min, students took the knowledge test within a maximum time limit of 40 min. To enhance student motivation we offered written performance feedback.

2.4. Analyses

All analyses of the manifest variables and factor score estimates were performed by means of SPSS 15. Amos 16.0 was used for structural equation analyses. First, we tested whether the measurement and theoretical model proposed by Beauducel et al. (2001) and Liepmann et al. (2007) fitted the data. Second, we tested for sex differences. Three approaches were used. First, we tested whether the sum scores differed with respect to sex, second, we investigated sex differences in the factor score estimates corresponding to the latent variables, and, third, we investigated possible sex differences using multiple indicator-multiple cause (MIMIC) models with sex predicting the five latent factors representing verbal, numerical, figural, fluid and crystallized intelligence. For all analyses, we applied a significance level of $p \leq .01$. For the MIMIC models, we applied the same procedure as depicted in Keith et al. (2008, pp. 507–509).

Prior to the MIMIC analyses we conducted preliminary analyses testing whether the prerequisites of the analyses were fulfilled. To this end we tested whether the variance–covariance matrices and the measurement models for males and females were equivalent using multi-group confirmatory factor analysis. Additionally to the chi-square statistics, model fit was tested by the comparative fit index (*CFI*), the root mean square error of approximation (*RMSEA*), and the standardized root mean square residual (*SRMR*), as proposed by Beauducel and Wittmann (2005). Furthermore, we report the non-normed fit index (*NNFI*). The following cut-off scores for these indices were used: $NNFI \geq .95$; $CFI \geq .95$, $RMSEA \leq .06$, $SRMR \leq .09$ (Hu & Bentler 1998, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003).

With respect to the choice of factor score estimates, Lawley and Maxwell (1971) took the position that no general preference for any factor score estimator could be given and that the required properties of the estimators were important. In the present context, the factor score estimates should have the same intercorrelations as the latent variables. If the intercorrelations between the factor score estimates were different from the intercorrelations of the latent variables (which could occur with some type of factor score estimates), some sex differences that might have occurred on one latent variable might partly occur on another factor score estimate. To avoid this problem, McDonald's (1981) correlation preserving factor score estimate was used in this study.

3. Results

First, analyses for the manifest variables are presented. Means and standard deviations of manifest sum scores of the subtests and broad cognitive abilities are separately presented for males and females in Table 1. For each score we performed analysis of variance (ANOVA) to test for sex differences in the manifest variables.

Males scored higher on all subtests indicating numerical intelligence and knowledge, on verbal analogies and two figural subtasks ("Abstract Pieces and Cubes"). At the scale level, males performed better on all assessed broad cognitive abilities.

Second, we performed structural equation modelling. As a first step, the fit of the present data to the theoretical model proposed by Liepmann et al. (2007) and Beauducel et al. (2001) was tested (see Fig. 1). The fit of the data was excellent ($CMIN = 190.52$, $df = 113$, $p < .001$; $NNFI = .97$; $CFI = .98$; $RMSEA = .03$; $SRMR = .03$). The model as well as the estimated parameters is presented in Fig. 1.

Based on the loadings and interfactor correlations of this model, we calculated McDonald's factor score estimates. Table 2 presents the means and standard deviations of the factor score estimates separately for males and females as well as the tests for group differences. We tested for sex differences in the factor score estimates by means of ANOVAs. Males scored significantly higher on the factor score estimates for fluid and crystallized intelligence as well as on numerical and figural intelligence. Females scored significantly higher on verbal intelligence.

Second, we checked for equivalence of the variance–covariance matrices for males and females. If the variance–covariance matrices are equivalent in the groups of males and females, the MIMIC analyses will produce the same results as multi-group analysis of mean and covariance structures (cf. Muthén, 1989; Reynolds et al., 2008). The latter approach is a further common statistical approach to compare latent means between groups such as sexes (e.g., van der Sluis et al., 2006, 2008). As the MIMIC model tests fewer parameters, i.e. the tested models are more parsimonious, and it is easier to interpret, we decided on this approach. The model specifying the equality of the variance–covariance matrices across sexes fitted the data extremely well ($CMIN = 267.42$, $df = 226$, $p = .03$; $NNFI = .98$; $CFI = .99$, $RMSEA = .01$; $SRMR = .03$).

In addition, the measurement invariance of the model was investigated by means of multiple-group analysis of mean

Table 1
Means (*M*) and standard deviations (*SD*) of students' manifest sum scores separately presented for males ($n = 426$) and females ($n = 551$) as well as ANOVA results and effect sizes for sex differences.

Cognitive ability	Females		Males		ANOVA		
	M	*SD*	*M*	*SD*	$F_{(1,976)}$	*p*	*d*
Sentence completion	12.91	2.67	13.01	2.86	.34	.56	−.04
Verbal analogies	11.03	2.55	12.15	2.56	46.07	<.001	−.44
Verbal similarities	10.97	3.17	11.34	3.04	3.44	.06	−.12
Calculations	12.08	3.32	13.89	3.41	69.79	<.001	−.54
Number series	13.52	4.25	15.51	3.98	55.33	<.001	−.48
Signs	12.01	3.61	14.83	3.74	142.07	<.001	−.77
Abstract pieces	10.61	3.57	11.84	3.68	27.82	<.001	−.34
Cubes	10.94	3.67	12.20	3.84	27.17	<.001	−.34
Matrices	11.39	2.61	11.06	2.90	3.49	.06	.12
Verbal knowledge	14.61	3.27	16.50	3.29	79.63	<.001	−.58
Numerical knowledge	13.45	2.88	15.31	3.33	87.22	<.001	−.60
Figural knowledge	15.31	3.46	17.76	3.63	115.23	<.001	−.69
Verbal intelligence	49.51	8.20	52.99	8.03	44.13	<.001	−.43
Numerical intelligence	51.07	10.20	59.54	10.79	157.59	<.001	−.81
Figural intelligence	48.26	8.88	52.87	9.41	64.46	<.001	−.50
Fluid intelligence	105.46	16.71	115.83	16.63	92.91	<.001	−.62
Crystallized intelligence	43.37	7.59	49.57	8.39	146.11	<.001	−.78
Overall cognitive ability	148.84	21.39	165.40	21.90	141.11	<.001	−.77

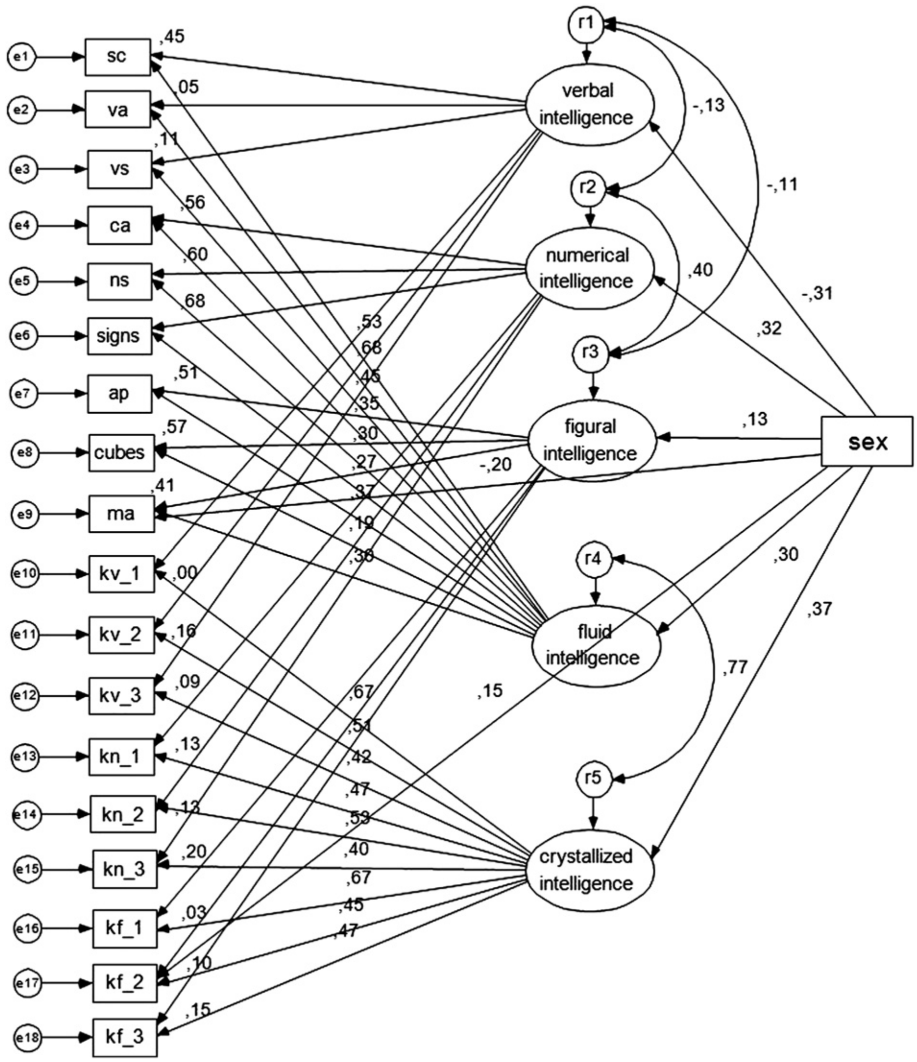

Fig. 1. Intelligence model as proposed by Liepmann et al. (2007) and Beauducel et al. (2001). Content factor path weights are presented on the left side, path weights from fluid and crystallized intelligence are shown on the ride sight. Sc = sentence completion, va = verbal analogies, vs = verbal similarities, ca = calculation, ns = number series, ap = abstract pieces, ma = matrices, kv = verbal knowledge, kn = numerical knowledge, and kf = figural knowledge.

and covariance structures. The measurement model did not differ significantly between men and women as the comparison between the constrained and unconstrained model showed ($CMIN = 49.49$, $df = 32$, $p > .01$). However, when additionally constraining intercepts, we found a significant decrease in model fit ($CMIN = 274.68$, $df = 18$, $p < .001$). This result was confirmed by the MIMIC model. The MIMIC model had a sufficient but not a very good model fit ($CMIN = 319.27$, $df = 127$, $p < .001$; $NNFI = .93$; $CFI = .95$; $RMSEA = .04$, $SRMR = .03$). In a MIMIC model, measurement invariance can be tested for the thresholds identifying those subtests that cause the measurement invariance. A significant direct effect of the covariates on the factor indicators signifies measurement non-invariance. To check whether there was any variance specific to a subtest which was related to sex, modification indices were checked. The largest MI suggested

Table 2
Means, standard deviations of the factor score estimates for females ($N = 551$) and males ($N = 426$) as well as ANOVA results and effect sizes for sex differences.

Cognitive ability	Females		Males		ANOVA		
	M	*SD*	*M*	*SD*	$F_{(1,976)}$	*p*	*d*
Verbal intelligence	0.10	0.97	−0.13	1.02	13.10	<.001	.23
Numerical intelligence	−0.21	0.96	0.27	0.99	57.96	<.001	−.49
Figural intelligence	−0.08	0.97	0.10	1.03	8.33	.004	−.19
Fluid intelligence	−0.26	0.94	0.34	0.98	93.41	<.001	−.62
Crystallized intelligence	−0.31	0.88	0.41	1.00	142.39	<.001	−.77

a direct path from sex to the subtest “Matrices” (MI = 30.33). After adding this path, MIs suggested another direct path from sex to the second figural knowledge item parcel (kf_2) (MI = 15.36). No further adjustments were necessary. The changes in the MIMIC model lead to an improved model fit (*CMIN* = 258.30; *df* = 125, *p* < .001; *NNFI* = .95; *CFI* = .97; *RMSEA* = .03; *SRMR* = .03). The effects of sex on the latent intelligence factors were not altered by these adjustments. This result was additionally supported by a further multi-group analysis. We fixed the factor loadings, the interfactor correlations, the error terms and the intercepts of the observed variables to be equal for males and females. The model had an acceptable fit (*CMIN* = 624.00, *df* = 302, *p* < .001; *NNFI* = .95; *CFI* = .95, *RMSEA* = .05; *SRMR* = .07). The same sex differences in the latent intelligence factors were demonstrated in this model as shown by the MIMIC model. Because loadings, interfactor correlations, error terms, and intercepts were equal for males and females, the fit of the multiple-group model indicated that the means of males and females on the latent variables were comparable. For reasons of simplicity and because in a MIMIC model it is possible to control for partial measurement invariance (cf. Keith et al., 2008), we report the detailed statistics for the MIMIC model in the following. The adjusted MIMIC model and the parameter estimates are presented in Fig. 2.

Next, we compared the effect sizes of the calculated sex differences based on the three different methodological approaches. As effect sizes of sex differences in latent cognitive abilities, we chose correlations between sex and the specific intelligences as estimated by the different methods as well as Cohen's *d*. Correlation coefficients were compared by performing Fisher *z*-tests for independent samples. Table 3 presents the effect sizes as well as the results of the pair-wise comparisons between the correlation coefficients.

Magnitude and direction of the found effect sizes for crystallized and fluid intelligence did not differ by method. However, magnitude and direction of the effect sizes found

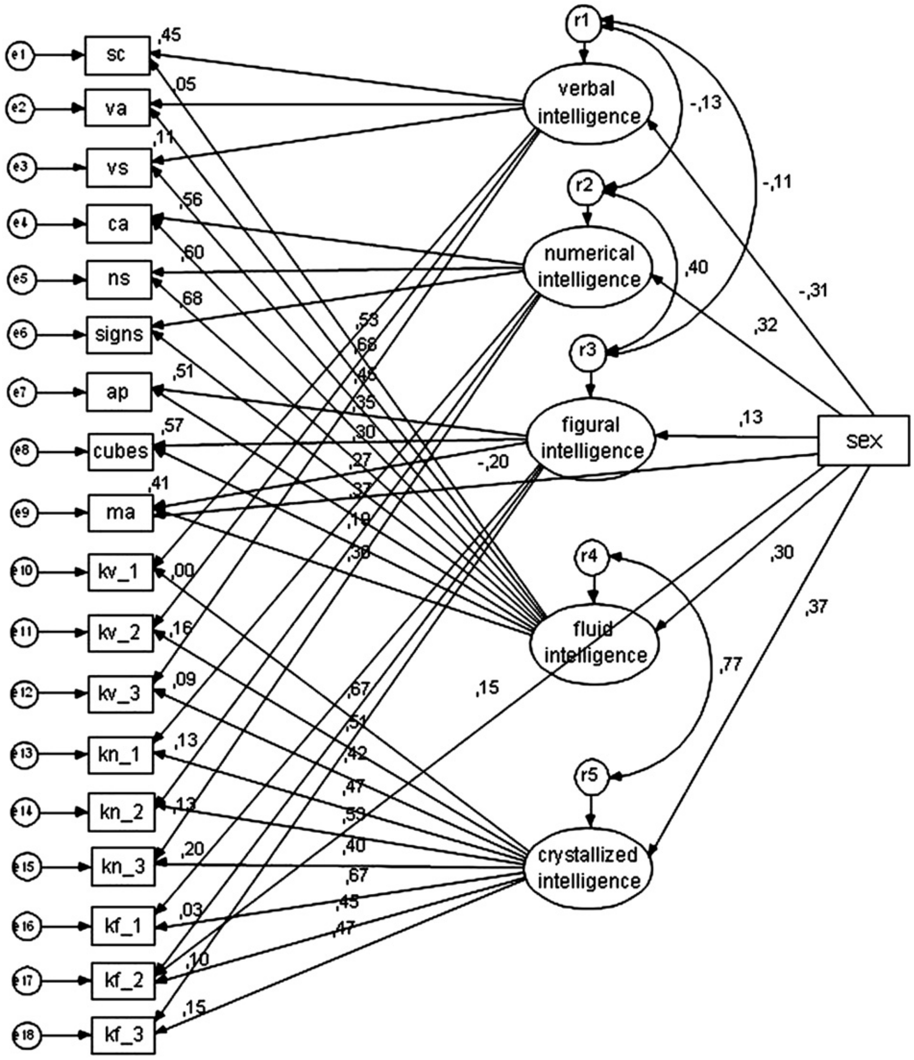

Fig. 2. MIMIC intelligence model with sex as a predictor. Content factor path weights are presented on the left side, path weights from fluid and crystallized intelligence are shown on the ride sight. Sc = sentence completion, va = verbal analogies, vs = verbal similarities, ca = calculation, ns = number series, ap = abstract pieces, ma = matrices, kv = verbal knowledge, kn = numerical knowledge, and kf = figural knowledge.

Table 3
Effect sizes of sex differences expressed as correlations between sex and specific intelligence measures based on group comparisons of manifest sum scores, factor score estimates, and latent variables.

Cognitive ability	Methodological approach						Correlation comparison		
	1) Manifest sum scores		2) Factor score estimates		3) Latent variables		*z*-value 1)–2)	*z*-value 1)–3)	*z*-value 2)–3)
	r	*d*	*r*	*d*	*r*	*d*			
Verbal	.21	.26	−.12	−.22	−.34	−.72	7.37***	12.77***	5.40***
Numerical	.37	.74	.24	.55	.32	.68	3.17**	1.25	−1.92
Figural	.24	.29	.09	.23	.13	.32	3.41**	2.72**	−.89
Reasoning/fluid intelligence	.30	.62	.30	.62	.30	.63	.00	.00	.00
Knowledge/crystallized intelligence	.36	.78	.36	.77	.40	.87	.00	−1.03	−1.03

Note. Positive effect sizes indicate a male advantage.
** $p<.01$.
*** $p<.001$.

for sex differences in verbal intelligence significantly differed by the method applied. Concerning numerical and figural intelligence, direction did not differ by the method applied but magnitude did. These results indicate that the methodological approach to investigate sex differences (manifest sum scores, factor score estimates, MIMIC model) might affect size and direction of the found sex differences in some intelligence facets.

4. Discussion

The present study aimed at investigating whether the chosen methodological approach to investigate sex differences might lead to different results. To this end we compared women's and men's performance by means of manifest sum scores, factor score estimates and latent variables. Men scored higher on fluid and crystallized intelligence than women. Size and direction of effects on these factors were not influenced by the method applied. However, sex differences in verbal, numerical, and figural content of intelligence greatly varied in magnitude and direction depending on the methodological approach applied.

Before discussing these results some limitations of the present study must be mentioned as they are important for correctly interpreting the results. First, the investigated sample was not representative of the full population of the 16–18 year old adolescents. Because students attended the most demanding type of school in the tracked German school system (Gymnasium), the sample represents the higher end of the ability distribution. Furthermore, more females than males attend this kind of school. These aspects have several implications. First, because admission mainly depends on grades in elementary school, males attending this kind of school should be more intelligent than females to the extent that females tend to get better grades and mature more rapidly. Thus, the present results do not give insight into sex differences in intelligence in general but only in this specific population.

Second, the fact that the sample was preselected by school achievement and, therefore, intelligence leads to restricted variance affecting the covariance structure of the investigated scales. Thus, the covariance structure and the found results apply only to students at the high end of the intelligence spectrum. Further studies should investigate samples covering a more representative intelligence range.

Although school-based samples such as the present one have their limitations, they also have some merits. The probability to investigate representative samples within given school (in tracked school systems representative samples of the student population attending the specific kind of school) is high, as students might feel obliged to participate in a testing (Dykiert et al., 2009). Other recruitment techniques might result in self-selection related to intelligence and sex. For example, Hunt and Madhyastha (2008) provided support that sex and intelligence are related to the decision to take part in testing when participation is not mandatory. The probability for average intelligent females to take the SAT, which is not mandatory, is higher than for males of comparable intelligence. Taken together, investigating school samples seems to be a good strategy to gain insights into sex differences in intelligence, though the present sample was only representative of the student population attending a Gymnasium. These limitations should be kept in mind when reading the rest of the discussion.

As already demonstrated in other studies employing the IST 2000-R (e.g., Liepmann et al., 2007; Steinmayr & Amelang, 2006), we found that men scored higher on the knowledge test than women. This result is comparable to other studies' findings, which consistently found a male advantage in knowledge tests when applying different instruments (e.g., Ackerman et al., 2001; Lynn et al., 2002). Consequently, it seems that men outperform women in crystallized intelligence if it is measured via a knowledge test. This result seems to be irrespective of the items' content, i.e. whether contents were presented in a verbal, numerical, or figural form, because sex differences did not change after removing variance specific to content factors and measurement error. The latter has already been demonstrated by Lynn and Irwing (2002) who showed a male advantage in a latent factor representing knowledge when controlling for measurement error. Other authors operationalized crystallized intelligence differently and found diverging results. For example, in the study by Kaufman et al. (1995), crystallized intelligence was operationalized by means of reading competencies. As reported above, the authors found no sex difference. Because females often exceed males in reading competencies (OECD, 2007) or no sex differences are observed in this ability (cf. Hyde, 2005), the results by Kaufman et al. are in line with other studies. Consequently, sex differences in crystallized intelligence might be independent from the applied methodological approach investigated but depend on the different operationalizations of crystallized intelligence.

In the present study, men also scored higher on fluid intelligence. This is in line with IST 2000-R results for an adult sample ranging between 27 and 42 years of age (Steinmayr & Amelang, 2006). Thus, there seems to be an advantage of men between 16 and 40 years in the fluid intelligence score as measured by the IST 2000-R. However, this result cannot be generalized to fluid intelligence per se. Horn (1988, p. 660) depicted fluid intelligence as a cognitive ability primarily measured by tasks measuring reasoning. Raven's Standard Progressive Matrices, Cattell's Culture Fair Test, and the Primary Mental Abilities Inductive Reasoning Test are further tests developed to assess fluid intelligence. They all assess fluid intelligence via figural material. Lynn and Irwing (2004b) found males to consistently obtain higher scores on Raven's progressive matrices in a meta-analysis. This result was replicated by Colom and García-López (2002) in a large high school sample covering more than 4000 students. In the same sample the authors also applied the Primary Mental Abilities Inductive Reasoning Test and the Culture Fair Test. Women exceeded men on the Primary Mental Abilities Inductive Reasoning Test and no sex differences were found in the Culture Fair Test. Given that they found no systematic sex differences, the authors concluded that there is no gender effect in fluid intelligence and that inconsistencies in studies reporting such differences can be attributed to the tests used. This result must also be considered when interpreting sex differences in fluid intelligence assessed via the IST 2000-R.

Considering verbal, numerical, and figural intelligence, the methods we applied in the present study did impact on the found effect sizes. Whereas men scored higher in verbal manifest sum scores, women outperformed men in verbal factor scores and the verbal latent factor. This result is in line with the finding by Hyde and Linn (1988). The authors found women to score higher in most verbal tasks whereas men scored higher in verbal reasoning tests. Two out of three verbal subtests ("Verbal Analogies" and "Verbal Comparison") of the basic module of the IST 2000-R are measures of verbal reasoning. After removing variance due to crystallized and fluid intelligence, which is very close to reasoning (cf. Horn, 1988), from the verbal intelligence factor, the expected female advantage was demonstrated. Thus, the male advantage in verbal manifest sum scores might be explained by the high loading of reasoning on the verbal scales.

Effect sizes in numerical and figural intelligence were smaller in factor scores and latent variables than in manifest sum scores. Sex differences in figural intelligence nearly vanished when investigated by factor score estimates or latent variables. This might also be explained by removing variance due to fluid and crystallized intelligence. This result is interesting with regard to the following aspect. First, figural intelligence is often considered as a proxy of fluid intelligence (Schulze et al., 2005) and men's advantage in figural tasks is often interpreted as a sign of their superior fluid intellectual capacity. The present result shows that men's advantage in figural material might largely be attributed to their higher performance in fluid and crystallized intelligence as measured with the IST 2000-R. Men and women might not differ in their general handling of figural material.

However, the results concerning the content intelligence factors contradict the findings by Johnson and Bouchard (2007). The authors found larger effect sizes for sex differences in specific intelligence abilities after removing variance due to the *g* factor. The different results might be explained by the fact that the authors applied a different theoretical intelligence model (a hierarchical four-stratum intelligence model) and investigated by far more intelligence subtests measuring more intelligence areas (including perceptual speed and memory) than in the present study. On the other hand, the fact that in the present study some sex differences were found for the manifest sum scores that were not found for the latent variable model and the factor score estimates also indicates that some sex differences might be related to rather specific performance variance.

The fact that Johnson and Bouchard (2007) found different results investigating a different set of abilities hints at a further limitation of the present study. We investigated a specific set of cognitive abilities and three methods commonly used to generate intelligence scores. Thus, the present results gave insights on how the methods investigated affected sex differences when focusing these abilities and this intelligence model. However, whether the same diverging results are found when investigating a different set of abilities, like Johnson and Bouchard (2007) did, or further methods cannot be concluded from the present study. Further studies should also compare different methods when investigating sex or other group differences when focusing different sets of abilities or different intelligence models.

References

Ackerman, P. L., Bowen, K. R., Beier, M., & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology*, *93*, 797–825.

Aluja-Fabregat, A., Colom, R., Abad, F., & Juan-Espinosa, M. (2000). Sex differences in general intelligence defined as *g* among young adolescents. *Personality and Individual Differences*, *28*, 813–820.

Beauducel, A., Brocke, B., & Liepmann, D. (2001). Perspectives on fluid and crystallized intelligence: Facets for verbal, numerical, and figural intelligence. *Personality and Individual Differences*, *30*, 977–994.

Beauducel, A., Liepmann, D., Felfe, J., & Nettelnstroth, W. (2007). The impact of different measurement models for fluid and crystallized intelligence on the correlation with personality traits. *European Journal of Psychological Assessment*, *23*, 71–78.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*, 41–75.

Brody, N. (1992). *Intelligence.* San Diego: Academic Press.

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, *32*, 445–451.

Colom, R., García, L. F., Juan-Espinosa, M., & Abad, F. J. (2002). Null sex differences in general intelligence: Evidence from the WAIS-III. *The Spanish Journal of Psychology*, *5*, 29–35.

Deary, I. J., Thorpe, G., Wilson, V., Starr, J. M., & Whalley, L. J. (2003). Population sex differences in IQ at age 11: The Scottish mental survey 1932. *Intelligence*, *31*, 533–542.

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

Dykiert, D., Gale, C. R., & Deary, I. J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence*, *37*, 42–47.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, *43*, 95–103.

Halpern, D. F., & LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review*, *12*, 229–246.

Horn, J. (1988). Thinking about human abilities. In J. R. Nesselroade R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 645–685). 2nd ed. New York: Plenum.
Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.
Hunt, E., & Madhyastha, T. (2008). Recruitment modeling: An analysis and an application to the study of male–female differences in intelligence. *Intelligence*, *36*, 653–663.
Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592.
Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53–69.
Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, *96*, 505–524.
Jackson, D. N., & Rushton, J. P. (2006). Males have greater *g*: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence*, *34*, 479–486.
Johnson, W., & Bouchard, T. J. J. (2007). Sex differences in mental abilities: *g* masks the dimensions on which they lie. *Intelligence*, *35*, 23–39.
Kaufman, J. C., Chen, T. -H., & Kaufman, A. S. (1995). Ethnic group, education, and gender differences on six Horn abilities for adolescents and adults. *Journal of Psychoeducational Assessment*, *13*, 49–65.
Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence*, *36*, 502–525.
Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*, 2nd ed. London: Butterworth.
Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligence-Structure-Test 2000 R (extended Edition) [Intelligenz-Struktur-Test 2000 R (erweiterte Auflage)].* Göttingen: Hogrefe.
Lynn, R. (1992). Sex differences on the differential aptitude test in British and American adolescents. *Educational Psychology*, *12*, 101–106.
Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences*, *17*, 257–271.
Lynn, R. (1999). Sex differences in intelligence and brain size: A developmental theory. *Intelligence*, *27*, 1–12.
Lynn, R., Allik, J., Pullmann, H., & Laidra, K. (2004). Sex differences on the progressive matrices among adolescents: Some data from Estonia. *Personality and Individual Differences*, *36*, 1249–1255.
Lynn, R., & Irwing, P. (2002). Sex differences in general knowledge, semantic memory and reasoning ability. *British Journal of Psychology*, *93*, 545–556.
Lynn, R., & Irwing, P. (2004). Sex differences on the advanced progressive matrices in college students. *Personality and Individual Differences*, *37*, 219–223.
Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*, 481–498.
Lynn, R., & Irwing, P. (2008). Sex differences in mental arithmetic, digit span, and *g* defined as working memory capacity. *Intelligence*, *36*, 226–235.
Lynn, R., Irwing, P., & Cammock, T. (2002). Sex differences in general knowledge. *Intelligence*, *30*, 27–39.
Maitland, S. B., Intrieri, R. C., Schaie, K. W., & Willis, S. L. (2000). Gender differences and changes in cognitive abilities across the adult life span. *Aging, Neuropsychology, and Cognition*, *7*, 32–53.
McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika*, *46*, 337–341.
Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.
Nyborg, H. (2003). Sex differences in *g*. In H. Nyborg (Ed.), *The scientific study of general intelligenceTribute to Arthur R. Jensen.* (pp. 187–222) Amsterdam: Elsevier Science Ltd.
Nyborg, H. (2005). Sex-related differences in general intelligence *g*, brain size, and social status. *Personality and Individual Differences*, *39*, 497–509.
Organization for Economic Co-Operation and Development (OECD). (2007). PISA 2006 Volume 2: Data Paris: OECD Publications.
Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, *36*, 236–260.
Rosén, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction*, *5*, 37–62.
Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.
Schulze, D., Beauducel, A., & Brocke, B. (2005). Semantically meaningful and abstract figural reasoning in the context of fluid and crystallized intelligence. *Intelligence*, *33*, 143–159.
Spearman, C. E. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.
Statistisches Bundesamt (Hrsg.). (2008). *Statistisches Jahrbuch 2008.* Für die Bundesrepublik Deutschland : Wiesbaden [Statistical year book 2008. For the Federal Republic of Germany].
Steinmayr, R., & Amelang, M. (2006). Erste Untersuchungen zur Kriteriums-Validität des I-S-T 2000 R an Erwachsenen beiderlei Geschlechts [First results regarding the criterion validity of the I-S-T 2000 R concerning adults of both sex]. *Diagnostica*, *52*, 181–188.
Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, *22*, 185–209.
Terman, L. M. (1916). *The measurement of intelligence.* Boston, MA: Houghton Mifflin.
van der Sluis, S., Derom, C., Thiery, E., Bartels, M., Polderman, T. J. C., Verhulst, F. C., et al. (2008). Sex differences on the WISC-R in Belgium and the Netherlands. *Intelligence*, *36*, 48–67.
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*, *34*, 273–289.
Wechsler, D. (1944). *Measurement of adult intelligence*, 3rd ed. Baltimore: William and Wilkins.