Available online at www.sciencedirect.com

ScienceDirect

Intelligence 36 (2008) 236–260

INTELLIGENCE

Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models ☆

Matthew R. Reynolds *, Timothy Z. Keith, Kristen P. Ridley, Puja G. Patel

The University of Texas at Austin, United States

Received 23 September 2006; received in revised form 18 May 2007; accepted 6 June 2007
Available online 17 July 2007

Abstract

Sex differences in the latent general and broad abilities underlying the Kaufman Assessment Battery for Children—Second Edition (KABC-II) were investigated for children and youth ages 6 through 18. The data were split into different age groups to account for changes due to differential development. Multi-group higher-order analysis of mean and covariance structures (MG-MACS) and multiple indicator-multiple cause (MIMIC) models were used to analyze these data. Boys consistently demonstrated a significant mean advantage on the latent visual–spatial ability (Gv) factor. A significant mean advantage was also found for boys on the latent crystallized ability (Gc) factor at all ages except for 17 and 18. Girls scored higher on the latent, higher-order *g* factor, at all ages, although this difference was statistically significant at only two age levels. An additional test, however, did not reveal a significant Age × Sex interaction effect, suggesting only main effects of Sex on Gv, Gc, and *g*.

Keywords: Sex differences; Intelligence; Multi-group confirmatory factor analysis; Broad cognitive abilities; MIMIC models; Three-stratum theory

1. Introduction

Sex differences in cognitive abilities have been a topic of debate among researchers for over a century. Researchers have focused on the study of differences in both general and specific cognitive abilities, and of differences in both means and variances. Some research reviews have found no meaningful differences in general intelligence between adult males and females, whereas others have concluded that adult males demonstrate a small to moderate advantage (cf. Halpern & LaMay, 2000; Irwing & Lynn, 2005; Jensen, 1998; Lynn & Irwing, 2004; Mackintosh, 1996; Nyborg, 2003). Some inconsistencies regarding sex differences in specific cognitive abilities are also present. Generally, males are considered to have an advantage in mathematical reasoning, mental rotation, and aspects of visual–spatial ability (Hulick, 1998; Jensen, 1998; Johnson & Bouchard, 2007; Voyer, Voyer, & Bryden, 1995), whereas females have generally shown advantages in numerical calculation, verbal fluency, and

☆ We are grateful to Alan Kaufman, Mark Daniel, and the American Guidance Service for providing the data used in this research. We owe a special thank you to Mark Daniel for graciously answering numerous questions about the test development process of the KABC-II. We also thank the three reviewers, Conor Dolan, Paul Irwing, and Wendy Johnson, for their thoughtful comments on earlier versions of this manuscript.

* Corresponding author. The University of Texas at Austin, Department of Educational Psychology, 1 University Station D5800, Austin, TX 78712-0383, United States. Tel.: +1 512 917 5184; fax: +1 512 475 7641.

E-mail address: matthew.reynolds@mail.utexas.edu
(M.R. Reynolds).

0160-2896/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2007.06.003

verbal memory (Jensen, 1998; Kimura, 2004; Maitland, Intrieri, Schaie & Willis, 2000).

Perhaps the most interesting recent studies of sex differences in psychometric IQ have investigated whether a general factor of intelligence (*g*), specific abilities (henceforth referred to as broad abilities), or both, explain differences in observed test scores. For instance, van der Sluis et al. (2006) and Dolan et al. (2006) concluded that sex differences in observed scores or first-order factors could not be explained by a second-order *g* factor. Rather, these authors concluded that males performed better on the broad ability factors of working memory and perceptual organization, and van der Sluis and colleagues concluded that females performed better on a processing speed factor. These two studies are of particular interest because the authors used multiple-group mean and covariance structure analysis so that sex differences could be attributed to the latent factors underlying the observed scores. Here, we investigate sex differences in cognitive abilities of children and adolescents using a similar methodological framework.

1.1. Sex differences in childhood and adolescence

1.1.1. Mean levels

In the present study, we investigated sex differences in mean levels of cognitive ability for children and adolescents aged 6 to 18. The existing evidence is mixed concerning the presence of significant sex differences in levels of general intelligence in children and adolescents. Although some research suggests no such differences (e.g., Camarata & Woodcock, 2006; Deary, Thorp, Wilson, Starr, & Whalley, 2003; van der Sluis et al., in press), other evidence suggests that differences do exist, but neither boys nor girls show a consistent advantage. For example, Arden and Plomin (2006) found that girls demonstrated a mean advantage on a *g* factor at ages 2, 3, 4, and 7. At age 9 there was no statistically significant difference, but by age 10, boys had surpassed girls. Alternatively, using hierarchical confirmatory factor analyses, Rosén (1995) found a girl advantage in a *G* (*G*, rather than *g*, is used to denote the general factor in nested factor models) factor for 12 and 13 year-olds and Härnqvist (1997) found a girl advantage in a *G* factor for 11 to 16 year-olds.

At minimum, these studies illustrate that developmental differences need to be considered when studying sex differences in children. For example, Rosén (1995) suggested that the girl advantage in *G* may have been due to an earlier onset of puberty and mental growth typically seen among girls. Therefore, it is important for cross-sectional studies to consider that, on average, girls mature physically and perhaps cognitively, faster than boys (Halpern, 1997; Lynn, 1999).

Studies with children and adolescents have shown some rather consistent sex differences in broad cognitive abilities. Boys have demonstrated advantages in a latent visual–spatial ability factor (Härnqvist, 1997; Rosén, 1995) and in tests of visual–spatial ability (Lynn, Fergusson, & Horwood, 2005). Alternatively, girls have shown higher levels of processing speed with even larger differences in middle and high school students (Camarata & Woodcock, 2006; Hulick, 1998; Lynn et al., 2005).

There have also been inconsistent findings. Findings of sex differences related to broad verbal ability and Gc have been conflicting. Although some studies have found girls outperforming boys on some measures of verbal ability (Hyde & Linn, 1988), a latent verbal factor (Härnqvist, 1997) and a latent Gc factor (Rosén, 1995), other studies have found boys outperforming girls on measures of vocabulary (Lynn et al., 2005), verbal ability (Rosén, 1995), and Gc composite scores (Camarata & Woodcock, 2006). Different findings may well be the result of the inconsistent operational definitions of Gc and verbal ability or by the use of measured variables in some studies versus latent variables in others.

1.1.2. Variability

Studies investigating sex differences in intelligence have focused primarily on means; less attention has been devoted to differences in variance across groups. Although there is a tendency for males to demonstrate greater variability in IQ or subtest scores, there is no conclusive evidence to support this notion (cf. Feingold, 1992; Hedges & Friedman, 1993; Nyborg, 2003). Little is known about variance differences in children and adolescents. Boys have been found to be more variable in *g* factor scores at ages 3, 4, 7, 9, and 10 (Arden & Plomin, 2006), in IQ scores at age 11 (Deary et al., 2003), but no differences have been found in a latent *g* factor from ages 11 to 16 (Härnqvist, 1997). In the context of broad abilities, boys have demonstrated more variability in a broad visual–spatial ability factor at ages 11, 12, and 13 (Härnqvist, 1997; Rosén, 1995), but girls have been shown to be more variable in Gv at ages 14, 15, and 16 (Härnqvist, 1997).

1.2. Methodological considerations

Debating the merits of the reviews or specific studies of sex differences is beyond the scope of this article.

However, one methodological issue highlighted in these reviews is worth reiterating: Researchers must be clear in terms of the ability constructs they are actually studying (Jensen, 1998; Nyborg, 2003). For example, comparing males and females on a global IQ score does not constitute a comparison of *g*, the construct of interest when studying general intelligence. A global IQ score is an index that provides information about the position of one person relative to others in a population. Although global IQ may hint at *g*, it does not represent the scientific construct of *g* which affects performance on a myriad of measures of cognitive tasks (Bartholomew, 2004). Instead, an IQ score is contaminated with measurement error, specific abilities, and others sources of non-*g* variance (Keith, 2006). Differences in global IQ scores may reflect differences in broad abilities, unique factors, or measurement error rather than differences in *g*. If two groups differ in a broad ability (e.g., visual–spatial ability) and a test is loaded with items that measure that ability, then these differences will likely show up as differences in a global IQ score (Bartholomew, 2004).

A similar concern pertains to less general cognitive abilities, such as Gc, Gv, and other broad abilities. Comparisons of broad abilities are often made on the basis of subtest or composite scores. These scores, however, are contaminated by *g*, specific abilities, and measurement error. The scores thus do not reflect the *theoretical constructs* of interest.

Comparisons of observed score variances are similarly problematic. Differences in observed score variances do not necessarily reflect variability differences in the *theoretical constructs* of interest, such as *g* or visual–spatial ability. To investigate whether variance and mean differences in ability constructs exist between males and females, observed test scores need to be decomposed first so that error and specific variance (and in broad ability factors, the variance attributed to *g*) are controlled. Before assertions are made that boys and girls differ in latent constructs underlying the test scores, factorial invariance must satisfied (Dolan, 2000; Widaman & Reise, 1997).

1.3. Latent variable modeling for studying sex differences

Human cognitive abilities have been conceptualized as latent constructs that underlie, among other things, performance on tests of cognitive ability. Three-stratum theory (Carroll, 1993), for example, posits three (or more) levels of cognitive abilities, with *g* affecting various broad abilities, and all those abilities affecting more specific abilities which affect test performance. Cognitive abilities are *not* conceptualized as emergent constructs caused by subtest variables or as linear combinations of test scores. Questions about group differences in cognitive abilities need to be answered within a latent variable system.

Analyses of mean differences in human cognitive abilities also need to account for the complex, hierarchical, and multi-factorial nature of intelligence (Keith, 2005). Modeling theoretical constructs within a latent variable framework allows for comparisons of theoretically error-free constructs and for the analysis of different abilities at different strata, or levels, simultaneously, and at the appropriate level. Latent variable modeling techniques, such as multi-group confirmatory factor analysis, also provide a flexible framework to test for factorial invariance across groups. These techniques are important because they evaluate whether the links between theoretical constructs and empirical observations are identical across groups. Simply put, “science requires invariance” (Meredith & Horn, 2001, p. 203). We are not the first to discuss the advantages and disadvantages of different methods for comparing group differences and refer the readers to other, more detailed sources (Dolan, 2000; Hancock, 1997; Lubke, Dolan, & Kelderman, 2001; Meredith, 1993).

In the present study, multi-group, higher-order analysis of mean and covariance structures (MG-MACS) and multiple indicator-multiple cause (MIMIC) analyses were used to examine latent mean and variance differences in cognitive abilities of children and adolescents ranging from 6 to 18 years in age. We performed empirical tests of whether the same constructs were measured across groups. We assessed whether mean and variance differences in subtest scores were due to 1) a second-order *g* factor, 2) broad ability factors, including crystallized ability or knowledge, fluid reasoning, visual–spatial thinking, long-term retrieval, and short-term memory, 3) both *g* and broad ability factors, or 4) unique aspects of subtests, include error and specificity.

The Kaufman Assessment Battery for Children—Second Edition (KABC-II) was used to test these differences; the KABC-II, designed to measure multiple, hierarchical cognitive abilities in children, was derived, in part, from Carroll's Three-Stratum Theory (1993)[1]. The KABC-II has been shown to have excellent psychometric properties (Kaufman & Kaufman, 2004;

[1] The test was grounded in Cattell–Horn–Carroll theory (a combination of Carroll's Three-Stratum and Cattell and Horn's Gf–Gc theory) and Luria's neuropsychological model (cf. Carroll, 1993; McGrew, 2005).

Reynolds, Keith, Fine, Fisher, & Low, in press) and the theoretically derived intelligence test is especially appropriate for the CFA analyses used here.

1.4. Purpose

This research focused on two specific questions:

Do mean levels of *g* and of five latent broad abilities differ between boys and girls at different ages on the KABC-II?

Does the variability of *g* and of five latent broad abilities differ between boys and girls at different ages on the KABC-II?

2. Method

2.1. Instrument

The KABC-II is an individually administered measure of cognitive abilities designed for use with children and adolescents ages 3 to 18. The KABC-II was standardized on a nationally representative (U.S.) sample of 3025 children and adolescents, aged 3 to 18; the sample "mirrors 2001 U.S. Census data with respect to gender, ethnicity, parental education level, geographic region..., and educational and psychological classifications..." (Braden & Ouzts, 2005, p. 518). The scale is designed, in part, to measure five of the broad abilities from Three-Stratum Theory, as well as a higher-order general intelligence.

The KABC-II includes sixteen subtests and the extended battery takes about 100 min to administer. The school-age subtests are described briefly in Table 1. The scoring structure of the test requires the subtests be summed to scale scores representing five broad abilities: crystallized ability (Gc), visual–spatial ability (Gv), fluid reasoning (Gf), long-term storage and retrieval (Glr), and short-term memory (Gsm); and a global IQ score (named the Fluid-Crystallized Index). The scores are age-standardized and the test has been shown to be invariant in its measurement across its 3 to 18 year age range (Reynolds et al., in press). The tests administered vary to some degree for preschool versus older children; for that reason, only scores from children and adolescents ages 6 through 18 were used in these analyses.

2.2. Participants

The KABC-II standardization sample for ages 6 to 18 was used for all of the analyses; this is the same sample used by Reynolds and colleagues (in press) to assess the validity of the KABC-II's higher-order structure. The sample was divided into ages 6 to 8, 9 to 11, 12 to 14, 15 to 16, and 17 to 18 for all analyses to assess possible developmental changes in sex differences in children across the ages. The distribution of race/ethnicity and sex for ages 6 to 18 is shown in Table 2.

2.3. Preliminary analyses

Means and standard deviations, along with the results of tests of mean differences and effect sizes of the individual subtests for each age group, are presented in Appendix A. Note that there are relatively few sex

Table 1
Description of KABC-II subtests for children and youth ages 6 to 18

Subtest	Description
Riddles	Examiner describes aspects of an object or idea; the child points to the object or names it
Verbal Knowledge	Child points to a picture that best represents the meaning of a vocabulary word or the answer to a general information question
Expressive Vocabulary	Child names pictured objects
Rover	Child determines the most efficient route for a dog to find a bone on a grid. The route must take into account various obstacles
Triangles	A block design-type task using two-colored foam triangles
Block Counting	Child counts blocks in pictures; some blocks are clearly visible, others are implied or only partially visible
Gestalt Closure	Examiner presents incomplete black and white drawings; the child describes the implied object or action
Story Completion	Child selects the most logical pictures needed to complete an incomplete, pictorially-presented story
Pattern Reasoning	Child completes a series of stimuli by correctly identifying the missing item from several choices
Rebus	Examiner teaches the meaning of rebuses; the child reads a series of rebuses, which form a sentence or phrase
Rebus Delayed	Childs reads a series of rebuses 15–25 min after initial training
Atlantis	Examiner teaches names for cartoon fish and other underwater objects; the child points to the correct picture when the examiner subsequently names them
Atlantis Delayed	Child points to the Atlantis objects 15–25 min after initial training
Word Order	Examiner states object names, child touches pictures of the objects in the same order. Later items have an intervening interference task
Number Recall	Common digit recall task
Hand Movements	Examiner makes a series of hand motions which the child repeats

Table 2
Demographic characteristics for ages 6 to 18 of the normative sample for the KABC-II

Variable	*N*	*N* by age group				
	Total	6–8	9–11	12–14	15–16	17–18
Total sample	2375	600	600	600	300	275
Sex						
Boys	1186	304	301	302	146	142
Girls	1189	296	299	298	154	133
Race/ethnicity						
White	1475	367	371	373	185	179
Hispanic	420	114	109	104	48	45
African American	352	90	87	88	48	39
Other	128	29	33	35	19	12

differences in the observed scores at the $p<.01$ level. Using an alpha level of .01 to judge statistical significance, equality of the covariance matrices can be assumed in each age group. MANOVA results showed a main effect of sex in three age groups (i.e., 9 to 11, 12 to 14, and 15 to 16) suggesting a sex difference on a multivariate linear combination of the subtests.

Preliminary analyses in the SPSS Missing Values Analysis package indicated that the null hypothesis of observations missing completely at random could not be rejected ($p=1$, at each age level). Skew and kurtosis of the subtests were not a concern, as skew and kurtosis were within a −1 to 1 range in each age group for all subtests.

2.4. Analytic approach

Multi-group, higher-order analysis of mean and covariance structures (MG-MACS) and multiple indicator-multiple cause (MIMIC) analyses were performed with the Amos 5 program (Arbuckle, 2003). All analyses were performed with the raw data files of age-standardized scores. Full information maximum likelihood estimation procedures were used to handle missing data. Tests of factorial invariance included comparisons of nested models. The likelihood ratio test (i.e., $\Delta\chi^2$) was used to test for statistically significant differences between these models. $\Delta\chi^2$ has been found to control for Type I error and provides acceptable power when testing for factorial invariance in single-factor models (French & Finch, 2006). In addition, Akaike's Information Criterion (AIC; Akaike, 1987) was used to compare the fit of competing models. A lower AIC value is indicative of a better-fitting model. We also used the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI), but primarily to judge the fit of single models (Hu & Bentler, 1999). The adjusted RMSEA was used for multiple group models (Steiger, 1998). The adjustment requires that the RMSEA be multiplied by the square root of the number of groups.

2.5. Analytic rationale

2.5.1. Multi-group analysis of mean and covariance structures

Before tests of latent mean or variance differences can be performed, the measurement instrument must demonstrate factorial invariance (Little, 1997; Widaman & Reise, 1997). In higher-order models, additional forms of factorial invariance must be achieved for group comparisons of the second-order factors to be meaningful (Byrne & Stewart, 2006; Chen, Sousa, & West, 2005). For purposes of this study, factorial invariance was assessed via the application of parameter constraints, moving sequentially from *unconstrained* to more *constrained* models which were compared with $\Delta\chi^2$ and AIC values. Because of the multiple tests, the complexity of the models, and the number of parameters tested, a probability level of .01 was used for model comparisons related to invariance tests at the measurement level (Little, 1997). The model specifications are summarized in Table 3.

The first test, configural invariance, was of whether the factor pattern for boys and girls was similar (Horn & McArdle, 1992; Widaman & Reise, 1997). All estimated loadings were free to vary across groups (Steenkamp & Baumgartner, 1998). Although configural invariance requires that the factor pattern be similar for boys and girls, the model does not test whether the same constructs are being measured across groups. A test of equal first-order factor loadings (or metric invariance) involving the specification of identical cross-sex subtest loadings on the first-order factors was required to assess whether the first-order factors were the same for boys and girls (Table 3, Model 2).

Satisfying first-order factor loading equivalence was a necessary but not sufficient condition for testing differences in latent means (Widaman & Reise, 1997). To test latent means, the sufficient condition of equal units of measurement and origins of scale (i.e., equal first-order factor loadings and subtest intercepts) needed to be satisfied (Chen et al., 2005; Meredith, 1993; Steenkamp & Baumgartner, 1998). In our study, if first-order factor loading or subtest intercept equality was not satisfied, then subtest loadings or intercepts were examined to determine whether the misfit was due to relatively few parameter constraints (Byrne, Shavelson, & Muthén, 1989). Such tests for partial invariance were not used often in this study, and this topic is discussed more in the Results section.

Table 3
Model tests for factorial invariance and latent mean differences

Models	Model constraints
1. Configural (CF)	One loading per factor = one; All factor mean differences (diff) = zero
2. Equal first-order factor loadings (FOL)	CF model+ Estimated first-order factor loadings = across groups
3. Equal subtest intercepts (SI)	FOL model+ Subtest intercepts = across groups; First-order factor mean diff estimated (i.e., girls first-order means = zero; boys first-order means free)

4. Equal subtest residual VAR/COV (SR)	SI model+
	Subtest residual variance (VAR) = across groups;
	Atlantis and Atlantis Delayed covariance (COV) = across groups;
	Rebus and Rebus Delayed COV = across groups

5. Equal second-order factor loading (SOL)	SR model+ Second-order factor loadings = across groups
6. Equal first-order unique VAR (FOU)	SOL model+ First-order unique VAR = across groups
7. Equal *g* VAR (*g*VAR)	FOU model+ *g* VAR = across groups

8. *g* mean difference (*g* mean diff) (i.e., Equal first-order means)	First- and second-order loadings = across groups;
	Subtest intercepts = across groups;
	Subtest residual VAR/COV = across groups;
	First-order unique and *g* VAR = across groups;
	First-order mean diff = zero;
	Second-order mean diff (i.e. girls second-order *g* mean = zero; boys second-order *g* mean free)

9. *g* mean diff with some first-order means diff	*g* mean diff+ SOME first-order mean diff; SOME first-order mean diff = zero

If first-order factor loading and subtest intercept equality was satisfied, we continued with a test of subtest residual equivalence (i.e., strict invariance). This level of factorial invariance required that the first-order loadings, subtest intercepts, and subtest residual variances/covariances be equal (Table 3, Model 4). This level of invariance establishes strict invariance at the measurement level (Little, 1997). This level of measurement invariance needs to be satisfied if observed score variances are to be compared across groups.

Our substantive hypotheses were tested with second-order models. Like the first-order model, if the second-order factor was to be considered the same for boys and girls, the second-order factor loadings, that is, the regressions of the first-order factors on *g*, needed to be equivalent across the groups. If first-order loading, subtest intercept, subtest residual variance/covariance, and second-order factor loading equivalence was tenable, then the substantive questions addressed in this research could be answered by making quantitative comparisons of the means and variances of the latent constructs (Little, 1997). These questions were answered with the use of a less strict statistical criterion than the criterion used at the measurement level because relatively fewer parameters were being estimated (Little, 1997). We chose a probability level of .05 for these tests of statistical significance.

One test of substantive interest was whether the variance of the broad abilities differed across the groups. This test was especially interesting because in a higher-order model the influence of *g*, subtest specific variance, and measurement error are removed so that the factors more closely represent the “true” constructs of interest. To test whether boys and girls rely on a similar range of broad abilities while performing cognitive tasks, the first-order unique variances were constrained to be equal across the groups. These constraints were of course in addition to the cross-group constraints already imposed on the first- and second-order loadings, subtest residual variances/covariances, and the subtest intercepts. Following a test of first-order unique variance differences, all of the constraints from the first-order unique variance model were retained, and the *g* variance was constrained to be equal across groups to test whether the variance of *g* was equally homogenous for boys and girls (Table 3, Model 7).

The remaining tests involved testing latent mean differences of the second-order *g* factor and of the first-order broad ability factors. Although we generally discuss differences in latent means, with a higher-order model the differences in the means of the broad abilities are, strictly speaking, differences in the first-order intercepts (Byrne & Stewart, 2006).

The first test was whether a second-order *g* factor accounted for all of the sex differences in the observed test scores. This test was carried out by restricting all of the first-order latent mean differences to zero while allowing a difference in the latent mean of *g* (Table 3, Model 8). If the model with these imposed constraints fit worse, then mean sex differences in *g* could not adequately account for mean sex differences in the subtest scores. Subsequently, we also tested for differences in broad ability means by allowing for mean differences in the first-order factor means. We summarize these model specifications in Table 3.

2.5.2. MIMIC model

If the factor loadings, residual variances, and factor variances were equal across groups, then a multiple indicator-multiple cause (MIMIC) model with a dummy coded, exogenous, sex variable was used to test for sex differences. Because the sex variable is error free, the unstandardized paths from sex to the latent abilities should be the same as differences in the latent means from the MG-MACS approach. The similarities between the MIMIC and MG-MACS approaches are analogous to the similarities between ANOVA and regression. Regression models can be used to conduct ANOVAs by simply adding a dummy coded variable as a predictor in the regression model (Hancock, 1997).

The model chi-square and degrees of freedom between the two modeling approaches do differ, however, because the MIMIC model tests fewer parameters. MIMIC models are not as flexible as MG-MACS models in that they make the assumption of a homogenous population covariance matrix. The MIMIC models are useful, however, to explain population heterogeneity in latent means (Muthén, 1989).

We performed the MIMIC analyses for two reasons. First, MIMIC models may be easier to interpret for some. For example, in our higher-order model, direct paths from sex to latent variables were included. Thus, with a direct path from sex to *g*, the unstandardized path coefficient from sex to the broad abilities (first-order factors) represented the mean difference on that broad ability considering boys and girls were at the same level of *g*. The effect of sex on subtest performance was indirect through the factors, except in cases where it was necessary to also include a direct path from sex to a subtest. The path from sex to a subtest was necessary

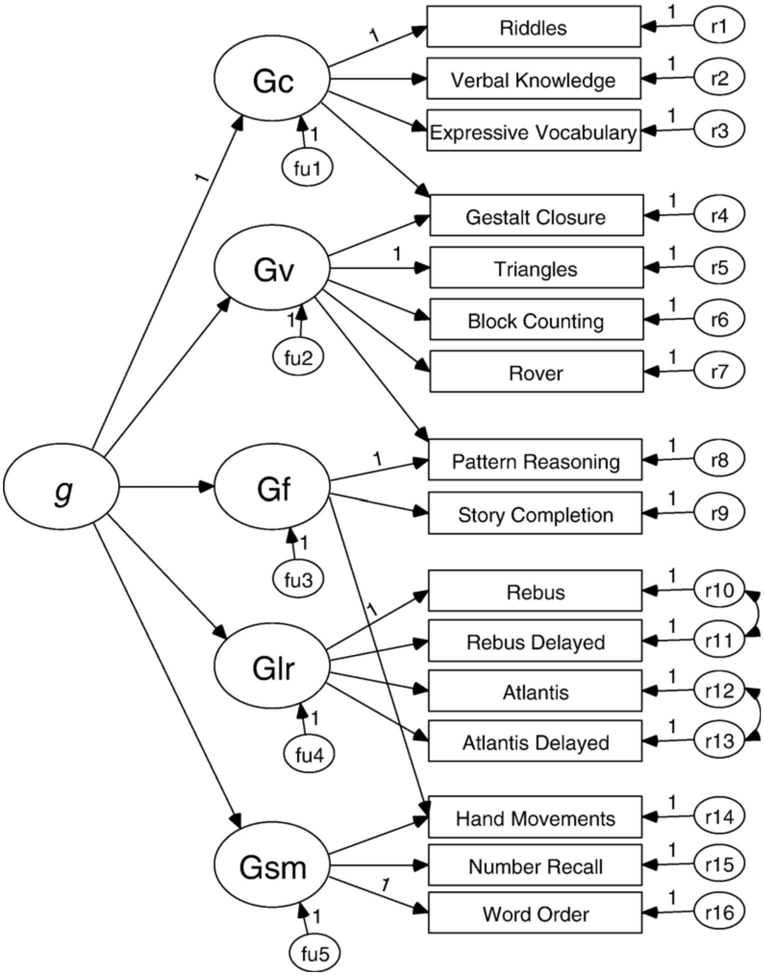

Fig. 1. The KABC-II factor structure.

when the factors could not account for all of the sex differences in a subtest. Allowing such a path is equivalent to allowing intercepts to vary on a subtest in the MG-MACS model.

Second, although the two approaches (MG-MACS and MIMIC models) are described as producing similar results given a homogenous population covariance matrix (Thompson & Green, 2006), there have been few applied studies that compare the two approaches. Thus, many researchers may not be aware of the similarities and differences between the two approaches. One study (Rosén, 1995) that did compare the two approaches when studying sex differences in cognitive ability found similar, but not identical, *t*-values for mean differences across the approaches. Rosén compared the MIMIC models with three multi-group models: a fully invariant multi-group model, partially invariant multi-group model, and a multi-group model with all parameters free to vary across groups. When testing for invariance, however, she found a few non-invariant factor variances (3 out of 13), error variances (2 out of 32), and factor loadings (3 out of 100). Therefore, slight differences in the fully-invariant multi-group model and the MIMIC model would be expected because the MIMIC model assumes that all of those parameters are invariant.

2.6. KABC-II factor structure

An example of a higher-order model that is similar to the models used for the analyses is shown in Fig. 1. The structure is identical to the best fitting structure of the KABC-II found in a previous study (Reynolds et al., in press). The structure consists of five first-order factors, Gc, Gv, Gf, Glr, and Gsm, and a second-order *g* factor. Gestalt Closure loads on Gv and Gc, Pattern Reasoning loads on Gv and Gf, and Hand Movements loads on Gf and Gsm. The untimed scores for Story Completion, Pattern Reasoning, and Triangles were used because previous research suggested their superiority (in model fit) over versions of the subtests that awarded time bonuses (Reynolds et al., in press).

3. Results

3.1. Ages 6 to 8

In the configural model, the factor pattern was specified to be the same across the groups. The factor loadings estimated within groups were free to vary across the groups. For model identification purposes, the first indicator of each first- and second-order factor was set to one. The intercepts of the subtest indicators were estimated freely across the groups while the factor means were set to zero. All latent variances were estimated freely across the groups. Information about the fit of this model is shown in Table 4. Although the model chi-squared was statistically significant, the fit of the configural model was excellent as indicated by the CFI and adjusted RMSEA (Hu & Bentler, 1999). Note the odd number of *df*. The Gf unique variance was fixed to zero for boys because it showed a small, but non-significant, negative variance. This constraint did not result in degradation of model fit (as judged by $\Delta\chi^2$ and AIC), and suggests that Gf is a perfect indicator of *g* in the boys sample and that the negative variance was likely related to sampling fluctuations. It should also be noted that the Gf unique

Table 4
Invariance model comparisons for 6 to 8 year olds

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	*p*	CFI	RMSEA	AIC
1. Configural	280.90	189				0.977	0.041	510.90
2. First-order factor loading	292.10	203	11.20	14	0.67	0.977	0.038	494.10
3. Subtest intercept	299.92	214	7.82	11	0.73	0.978	0.037	479.92
4. Subtest residual VAR/COV	322.54	232	22.62	18	0.21	0.977	0.037	466.54
5. Second-order loading	330.72	236	8.18	4	0.09	0.976	0.037	466.72
6. First-order unique VAR	334.32	241	3.60	5	0.61	0.976	0.035	460.32
7. *g* VAR	334.38	242	0.06	1	0.80	0.977	0.035	458.39
8. *g* mean diff (i.e. no diff in first-order means)	367.21	246	32.83	4	<0.01	0.969	0.041	483.21
9. *g*, Gv mean diff	344.97	245	22.24	1	<0.01	0.975	0.037	462.97
10. *g*, Gc, Gv mean diff	334.84	244	10.13	1	<0.01	0.977	0.035	454.84
11. Gc, Gv mean diff	340.11	245	5.27	1	0.02	0.976	0.035	458.11
12. MIMIC model	192.94	107				0.978	0.037	318.94
13. MIMIC with sex effect on *g* removed	198.27	108	5.33	1	0.02	0.977	0.037	322.27

Note. Compare models 2 to 11 with the previous model in the table. Compare model 13 with model 12.

variance was also not statistically significant (but positive) in the female sample. This finding is not uncommon in intelligence batteries and, historically, Gf is not considered to add much unique variance above and beyond the variance accounted for by *g* (e.g., Gustafsson, 1984; Keith, Fine, Reynolds, Taub, & Kranzler, 2006).

Equality of the first-order factor loadings was tested next. Cross-group equality constraints were imposed on all of the first-order factor loadings. This model did not fit statistically significantly worse than the configural model, and these loadings were constrained to be equal in all future tests (see Table 4, Model 1). Equality of the first-order factor loadings indicated that the unit of measurement was equal across the groups.

Next, a subtest intercept invariant model was specified to test the hypothesis of identical origins of the scale across the groups. In this model, all of the first-order factor loadings and subtest intercepts were constrained to be equal across the groups. Each first-order factor mean was fixed to zero for girls and estimated freely for boys. Girls served as the reference group for this model and all subsequent models. The estimated means for the boys represented the mean difference from girls. As shown in Table 4 (Model 3), the model with equal subtest intercepts did not differ significantly from the first-order factor loading invariant model (Model 2). Meaningful group comparisons were tenable (Widaman & Reise, 1997).

Equality of residual variances/covariances was tested next by retaining the first-order factor loading and subtest intercept constraints, and by imposing cross-group equality constraints on the subtest residual variances and covariances. These additional constraints did not result in degradation of model fit. *All* differences between girls and boys on the subtests were accounted for by the factors (Lubke et al., 2001; Widaman & Reise, 1997).

Next, because the KABC-II model is a second-order model consistent with three-stratum theory (Carroll, 1993), we tested for equality of the second-order factor loadings. While retaining the constraints from the equal residual variance/covariance model (i.e., Model 4 in Table 4), we also constrained the second-order factor loadings to be equal across the groups. As shown in Table 4 (Model 5), the fit of the model did not differ significantly from the equal residual variance/covariance model (Model 4). The second-order factor loadings were equivalent across the groups, and the structural relation of *g* to the first-order factors was the same for boys as it was for girls.

The remaining tests were of substantive interest to this study. First, the first-order unique variances were constrained to be equal across sex. These variances were of particular interest because each represented the “true” unique broad ability variance (Chen et al., 2005). The first-order unique variances were equal across the groups, suggesting that the Gc, Gv, Gf, Glr, and Gsm factor variances were equally homogenous across sexes (Table 4, Model 6). The constraints from this model were retained in subsequent models.

Equality in the variance of *g* was tested next by constraining the *g* variance to be equal across the sexes. Interestingly, a test of equality in *g* variances, with *g* modeled at the appropriate level, did *not* support the hypothesis of a statistically significant sex difference in *g* variance (Table 4, Model 6). This finding suggests that the dispersion of *g* is similar for boys and girls. A hypothesis of equal first- and second-order factor loadings, subtest intercepts, subtest residual variances/covariances, first-order unique variances, and second-order factor variance was tenable. Cross-group equality constraints were retained on all of these parameters when we tested for latent mean differences.[2]

The first test of latent mean differences was whether a second-order *g* factor difference in means was the sole source of mean differences between boys and girls. The first-order latent means were set to zero for boys and girls (Byrne & Stewart, 2006). To allow for mean differences in *g*, the mean of *g* was set to zero for girls, while the mean of *g* was estimated freely for boys. If the fit of this model was not significantly worse than Model 7, it would indicate that *g* accounts for *all* observed mean sex differences. As shown in Table 4, this model (Model 8) resulted in a significant deterioration of model fit: *g* could not account for all of the observed score mean differences between boys and girls.

Because sex differences in Gv means are often hypothesized (Rosén, 1995; Voyer et al., 1995), we specified another model that allowed for mean differences in Gv and *g*. The first-order Gv mean was estimated freely for boys, while the first-order Gv mean remained fixed at zero for girls. Like the previous model (Model 8), the mean of the second-order *g* factor was estimated for boys while the mean of the second-order *g*

[2] We also tested models that did not impose cross-group constraints on the factor variances. The difference between these models and the models with factor variances constrained was trivial. In particular, these models pointed to the same magnitude of mean differences across the sexes. For example, in the 6 to 8 age group, statistically significant mean differences were found *g*(Mdiff=−.459; SE=.201), Gv(Mdiff=.762; SE=.148), and Gc(Mdiff=.572; SE=.179).

factor was set to zero for girls. This model (Model 9), which allowed for mean differences in Gv in addition to *g*, fit significantly better than the model in which no mean difference in Gv was allowed (Model 8). The fit of this model (Model 9), although improved, still suggested that differences in *g* and Gv could not account for all of the mean sex differences.

Because a recent study found a boy advantage in Gc (Camarata & Woodcock, 2006), we estimated another model that tested for sex differences in Gc in addition to *g* and Gv. This model (Model 10) resulted in an improved model fit. Moreover, this model did not differ significantly from Model 7, suggesting that the sex differences in the subtests were adequately accounted for by the model. Although we do not report all of the models, we should note that mean differences were tested for all other first order factors (Glr, Gf, Gsm) as well (this procedure was followed at every age level). These parameters were tested in different orders (i.e., Gf freed first, Gsm freed first, Glr freed first), but none of these models suggested a different pattern of sex differences on first-order latent means. That is, Gv and Gc were the only broad abilities to show statistically significant mean sex differences.

The next step was to investigate the direction and the magnitude of the sex differences in latent means. With this "*g*, Gv, Gc mean difference" model (Model 10), the latent mean differences of *g* (Mdiff=-.45; SE=.20), Gv (Mdiff=.77; SE=.15), and Gc (Mdiff=.57; SE=.18) were all statistically significant at the $p<.05$ level. Because the girl mean was set to zero, positive values indicate a higher boy mean and negative values indicate a lower boy mean. Boys showed a higher level of Gv and Gc, whereas girls showed a higher level of *g*.

The final test was to constrain the difference in *g* to be zero across the groups. The mean of *g* was set to zero for boys (it was already set to zero for girls). Latent mean differences were allowed for Gv and Gc. As shown in Table 4 (Model 11), this model fit significantly

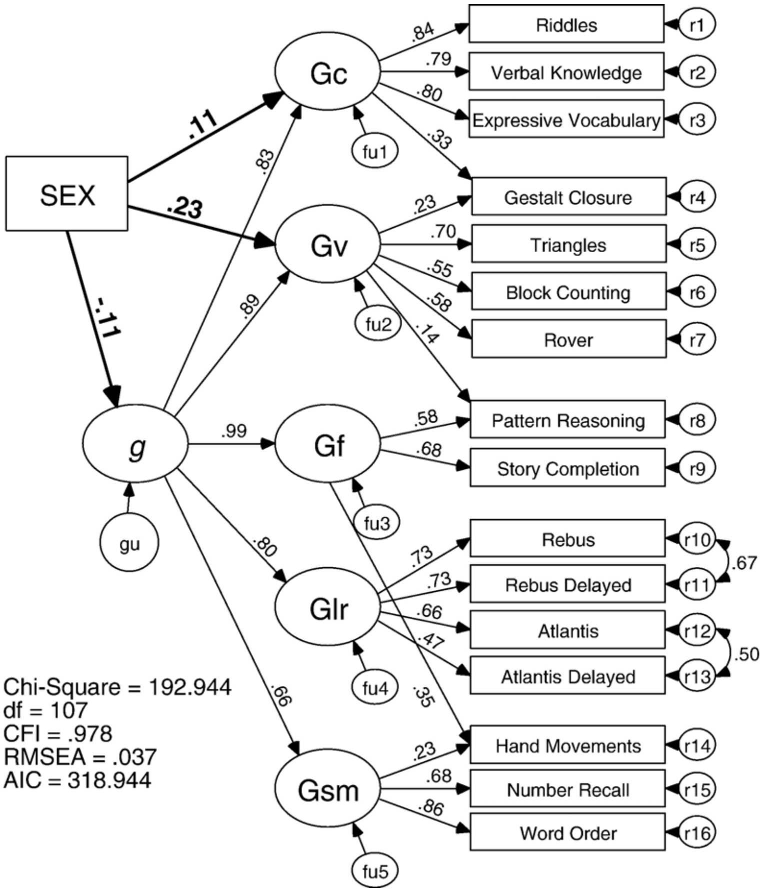

Fig. 2. The 6 to 8 year-old MIMIC model with statistically significant standardized effects of sex in bold and larger font. Negative paths from sex indicate a female advantage; positive paths from sex indicate a male advantage.

worse than did the model with the latent mean of *g* (Model 10) allowed to vary. This finding, like the previous comparison of the difference with its standard error, suggests that girls were statistically significantly higher on a latent *g* factor than were boys. Thus for this age group, all of the sex differences in the subtest scores were explained by *g*, Gv, and Gc.

Because the factor loadings, subtest residual variances/covariances, and factor variances were found to be invariant across sex, we also estimated a MIMIC model. For this analysis, a Sex variable was included to represent group membership (Girls = 0; Boys = 1). Direct paths were included from Sex to *g*, Gv, and Gc. This model represented a plausible model, as the fit statistics suggest excellent fit (Table 4, Model 12). Of particular interest, the unstandardized paths were identical to the mean differences obtained from the MG-MACS model. Thus, given the equality of variances and covariances (demonstrated via the MG-MACS model), the two models produced the same results.

The MIMIC model is depicted graphically in Fig. 2 with standardized paths coefficients and factor loadings. The path coefficient from Sex to *g* was negative and statistically significant, indicating a girl advantage, and the coefficients from Sex to Gv and Gc were positive and statistically significant, indicating a boy advantage in these two broad abilities. Deleting the path from Sex to *g* led to a statistically significant degradation in model fit, indicating that the effect of Sex on *g* is indeed statistically significant (Table 4, Model 13). These statistically significant paths from Model 12, which are shown in Fig. 2, can be interpreted as the direct effect of Sex on the latent second-order *g* factor and the effects of Sex on the latent first-order Gv and Gc factors after controlling for *g*. Hence, girls show higher average *g*, but when the level of *g* is held constant, boys show higher levels of Gv and Gc. Estimates of effect sizes can be obtained by multiplying the standardized effect by two (given that the SD of sex was .5). We show all estimated effects sizes converted to IQ points in Fig. 7.

3.2. Ages 9 to 11

The methods applied to the 6 to 8 sample data were applied to the data from the 9 to 11 age group. The configural model fit well (Table 5, Model 1). The first-order factor loadings were equal (Table 5, Model 2); however, not all of the subtest intercepts were equal (Model 3). To achieve partial intercept equality at the subtest level, the intercept of the Rover subtest was allowed to be estimated freely across groups while all of the other subtest intercepts were constrained to be equal (Model 4). This finding suggests that boys scored higher on Rover even after the effect of Gv (and the indirect effect of *g*) was considered. Similar to the 6 to 8 age group, as shown in Table 5, the hypotheses of equal subtest residual variances/covariances (Model 4), second-order loadings (Model 5), first-order factor unique variances (Model 6), and *g* variance (Model 7) were tenable. Likewise, mean differences in *g* alone could not completely account for sex differences in the observed scores (Table 5, Model 8). There were statistically significant differences in Gv (Mdiff = .72, SE = .17; Model 9) and Gc (Mdiff = .73; SE = .19; Model 10) with boys showing an advantage on both of these abilities. Although the mean of *g* was higher for girls (Mdiff = −.28; SE = .22), the difference was not statistically significant as indicated by both the standard error and the chi

Table 5
Invariance model comparisons for 9 to 11 year olds

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	*p*	CFI	RMSEA	AIC
1. Configural	270.72	188				0.979	0.038	520.72
2. First-order factor loading	296.03	202	25.31	14	0.03	0.976	0.040	500.03
3. Subtest intercept	322.17	213	26.14	11	<0.01	0.973	0.041	504.17
4. Subtest intercept but Rover diff	318.52	212	3.65	1	0.05	0.973	0.041	502.52
4. Subtest residual VAR/COV	340.55	230	22.03	18	0.23	0.972	0.040	488.55
5. Second-order loading	345.12	234	4.57	4	0.33	0.972	0.040	485.12
6. First-order unique VAR	352.41	239	7.29	5	0.20	0.972	0.040	482.41
7. *g* VAR	353.21	240	0.80	1	0.37	0.972	0.040	481.21
8. *g* mean diff (i.e. no diff in first-order means)	376.92	244	23.71	4	<0.01	0.967	0.042	496.92
9. *g*, Gv mean diff	366.90	243	10.02	1	<0.01	0.969	0.041	488.90
10. *g*, Gc, Gv mean diff	353.22	242	13.68	1	<0.01	0.972	0.040	477.22
11. Gc, Gv mean diff	354.85	243	1.63	1	0.20	0.972	0.040	476.85
12. MIMIC model	181.39	106				0.981	0.034	309.39

Note. Compare models 2 to 12 with the previous model in the table.

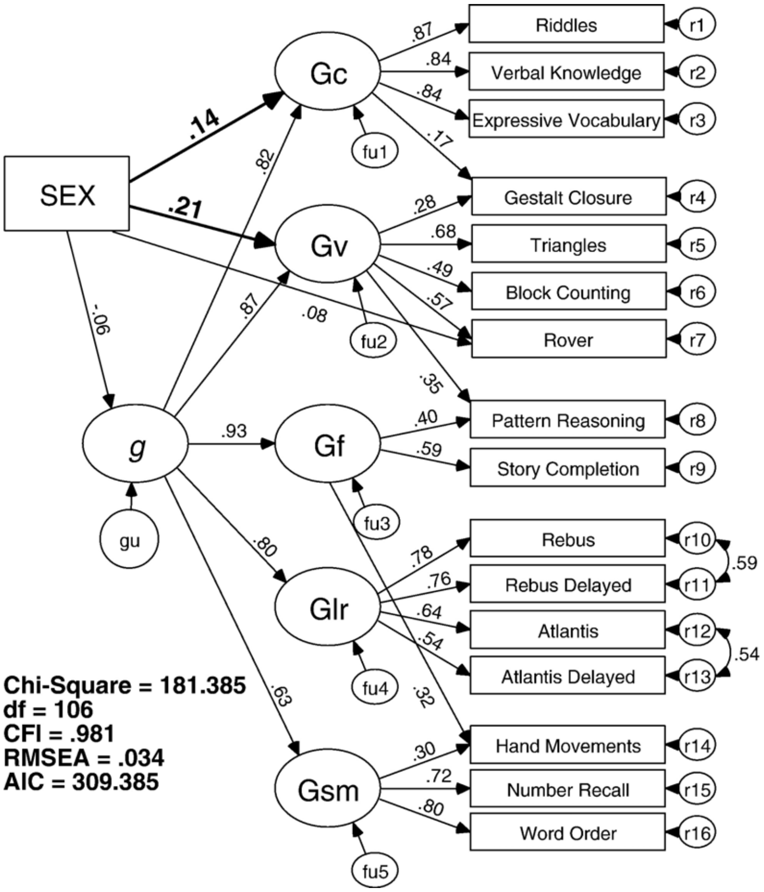

Fig. 3. The 9 to 11 year-old MIMIC model with statistically significant effects of sex in bold and larger font. Negative paths from sex indicate a female advantage; positive paths from sex indicate a male advantage.

square difference test (with differences in *g* constrained to be zero across the groups; Model 11). Hence, for 8 and 9 year-olds, when the intercept of Rover was free to differ between groups, Gv and Gc accounted for the sex differences in the observed scores. No other first-order factors showed sex differences.

Table 6
Invariance model comparisons for 12 to 14 year olds

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	*p*	CFI	RMSEA	AIC
1. Configural	286.03	189				0.979	0.041	516.03
2. First-order factor loading	295.08	203	9.05	14	0.83	0.981	0.040	497.08
3. Subtest intercept	319.08	214	24.00	11	0.01	0.978	0.041	499.08
4. Subtest residual VAR/COV	349.28	232	30.20	18	0.04	0.972	0.040	493.28
5. Second-order loading	353.86	236	4.58	4	0.33	0.975	0.041	489.86
6. First-order unique VAR	367.29	241	13.43	5	0.02	0.973	0.042	493.28
7. First-order unique VAR; Gsm VAR estimated freely	361.19	240	6.10	1	0.01	0.974	0.041	489.19
8. *g* VAR[a]	369.91	242	2.62	1	0.11	0.973	0.042	493.91
9. *g* mean diff (i.e. no diff in first-order means)	404.28	246	34.37	4	<0.01	0.967	0.047	520.28
10. *g*, Gv mean diff	385.45	245	18.83	1	<.0.01	0.970	0.044	503.45
11. *g*, Gc, Gv mean diff	373.82	244	11.63	1	<0.01	0.973	0.042	493.82
12. Gc, Gv mean diff	375.29	245	1.47	1	0.23	0.972	0.040	493.29
13. MIMIC model	194.65	107				0.982	0.037	320.65

Note. Compare models 2 to 7 and 9 to 12 with the previous model in the table; compare model 8 with Model 6.

[a] Although statistically significant, the variance of Gsm was constrained in subsequent models reported here. This constraint made no difference in the substantive findings.

For the MIMIC model, a Sex variable had direct paths to *g*, Gv, Gc, and the Rover subtest (to account for the finding of non-equivalent intercepts). The effects of Sex on Gv (β=.21, b=.72, SE=.17) and Gc (β=.14, b=.73, SE=.20) were statistically significant and the unstandardized loadings were identical to the mean differences obtained from the multiple group analysis. Boys demonstrated higher Gv after controlling for the level of *g*. Boys also scored higher on the Rover (β=.08, b=.45, SE=.23, p=.05) subtest even after accounting for the effects of Gv and *g*. Fig. 3 shows this model in path diagram form. Note that although *g* was not significant, we included the path to show consistent path models across the ages.

3.3. Ages 12 to 14

In the configural model, the variance of Gf was constrained to zero in the boy sample (Table 6, Model 1). As shown in Table 6, invariance tests of the first-order factor loadings (Model 2), subtest intercepts (Model 3), subtest residual variances/covariances (Model 4), and second-order factor loadings (Model 5) did not result in models that fit significantly worse. Constraining the five first-order factor unique variances to be equal did result in a model that fit worse, based on the AIC and $\Delta\chi^2$, using a $p<05$ criterion (Model 6). Gsm was more variable for girls than it was for boys (Model 7). Because it is possible that this finding may be related to an outlier(s) in the female data, we performed an additional check for multivariate outliers. No significant multivariate outliers were noted for the three Gsm subtests. Although the difference in Gsm variance was significant, not allowing it to be freely estimated across sex did not change any other substantive findings for this model. To be consistent with our models across age groups, in Table 6 we report only the models with the Gsm

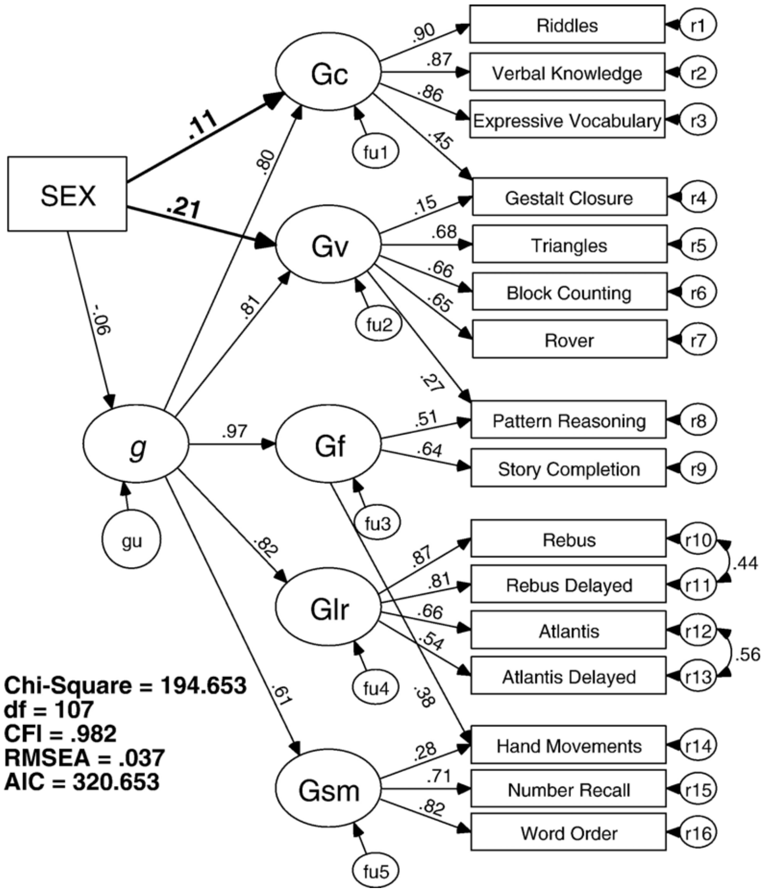

Fig. 4. The 12 to 14 year-old MIMIC model with statistically significant standardized effects of sex in bold and larger font. Negative paths from sex indicate a female advantage; positive paths from sex indicate a male advantage.

variance constrained to be equal. A model specified to test second-order factor variance equality (Model 8) did not fit worse than Model 6. Similar to the previous models at younger ages, the Gv (Model 9) and Gc (Model 10) first-order means were significantly different, but no other first-order mean differences were significant. The boys' Gv (Mdiff=.84; SE=.17) and Gc (Mdiff=.65; SE=.19) means were higher than the girls', both statistically significant. As shown in Table 6 (Model 10) boys and girls did not differ significantly in *g* (Mdiff=−.26; SE=.22) using either the standard error or $\Delta\chi^2$ criterion. From the MIMIC model we obtained the standardized and unstandardized effects of Sex on Gv (*β*=21, *b*=84) and Gc (*β*=.11, *b*=.65) controlling for the level of *g*.

The MIMIC model with standardized factor loadings and path estimates is shown in Fig. 4. Although the path from Sex to *g* was not statistically significant, the non-significant effect is included in the model for consistency with other reported models.

3.4. Ages 15 to 16

The configural model was a plausible model (Table 7, Model 1). Note, however, that the Gf unique variance for girls and boys was constrained to zero suggesting that Gf is a perfect indicator of *g*. All invariance tests were satisfied (Table 7, Models 1 to 5) and comparisons of latent means and variances were tenable. Again, the first-order factor unique variances (Model 6) and the *g* variance (Model 7) did not differ significantly between groups. Allowing for a mean difference in *g* resulted in a model in which all of the sex differences in observed scores were not accounted for (Model 8). Boys showed an advantage in Gv (Mdiff=1.11; SE=.21) and Gc (Mdiff=1.00; SE=.28) after controlling for *g*, and girls showed a statistically significant advantage in *g* (Mdiff=−.80; SE=.32). No other first-order factor means were significantly different.

Regression estimates obtained from the MIMIC model showed standardized effects for Sex of .17 on Gc (*b*=1.00), .31 on Gv (*b*=1.11), and −.17 (*b*=−.80) on *g*. Deleting the path from Sex to *g* led to a statistically significant degradation in model fit, indicating that the effect of Sex on *g* is indeed statistically significant (Table 7, Model 13). Again, positive values indicate that boys performed better whereas negative values indicate girls performed better. The estimates displayed in Fig. 5 are standardized. Note that the unique variance of Gf was set to zero in the MIMIC model.

3.5. Ages 17 to 18

The configural model for ages 17 to 18 was plausible. The Gf unique variance was set to zero in both groups. Equality of first- and second-order loadings, subtest residual variance/covariances, and subtest intercepts was tenable (see Table 8, Models 1–5). No group difference was found for the first-order unique variances or for the *g* variance. The only statistically significant difference in latent means was on the Gv factor (Mdiff=.96, SE=.24, *p*<.05). Mean differences in Gc (Mdiff=.44, SE=.26, *p*=.08) and *g* (Mdiff=−.54, SE=.30, *p*=.08) were in the same direction as with the other age groups, but were not statistically significant.

Table 7
Invariance model comparisons for 15 to 16 year olds

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	*p*	CFI	RMSEA	AIC
1. Configural	212.34	190				0.990	0.020	440.34
2. First-order factor loading	224.23	204	11.89	14	0.62	0.991	0.026	424.23
3. Subtest intercept	233.51	215	9.28	11	0.60	0.992	0.024	411.51
4. Subtest residual VAR/COV	254.09	233	20.58	18	0.30	0.989	0.024	396.09
5. Second-order loading	257.88	237	3.79	4	0.44	0.991	0.024	391.88
6. First-order unique VAR	262.19	241	4.31	4	0.37	0.991	0.024	388.19
7. *g* VAR	262.41	242	0.22	1	0.64	0.991	0.024	386.41
8. *g* mean diff (i.e. no diff in first-order means)	298.04	246	35.63	4	<0.01	0.978	0.038	414.04
9. *g*, Gv mean diff	278.67	245	19.37	1	<0.01	0.986	0.030	394.67
10. *g*, Gc, Gv mean diff	263.12	244	15.55	1	<0.01	0.992	0.023	383.12
11. Gc, Gv mean diff	269.74	245	6.62	1	0.01	0.989	0.026	387.74
12. MIMIC model	135.77	108				0.988	0.029	259.78
13. MIMIC with sex effect on *g* removed	142.42	109	6.65	1	<0.01	0.986	0.032	264.42

Note. Compare models 2 to 11 with the previous model in the table. Compare model 13 with model 12.

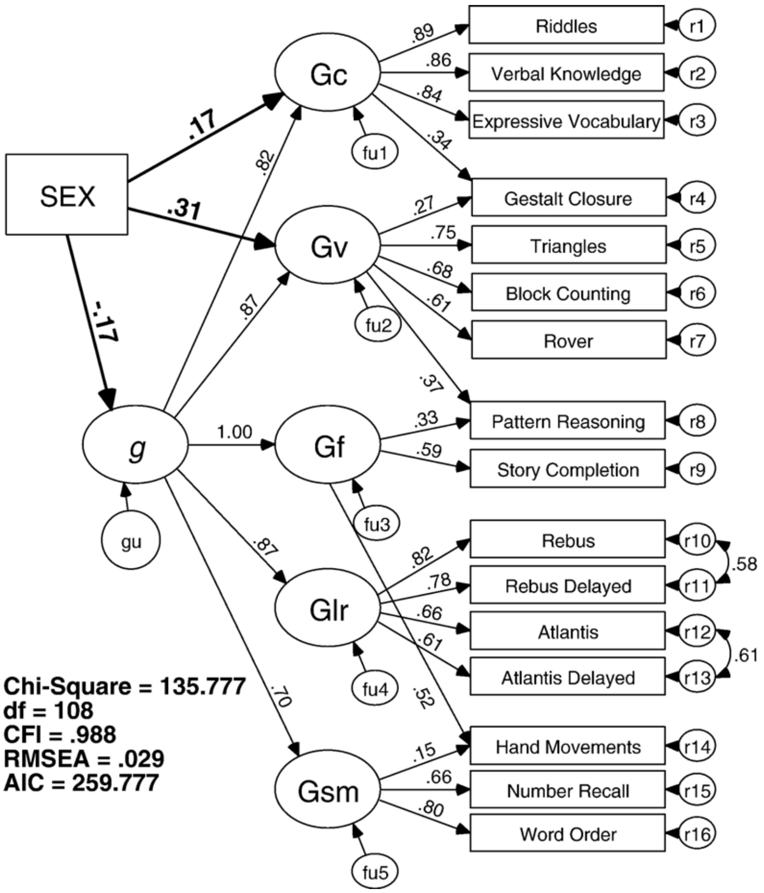

Fig. 5. The 15 to 16 year-old MIMIC model with statistically significant standardized effects of sex in bold and larger font. Negative paths from sex indicate a female advantage; positive paths from sex indicate a male advantage.

Again the MIMIC model resulted in the same unstandardized path of Sex to Gv (b=.96). The standardized estimate was .22. The MIMIC model with the effects of Sex on *g*, Gv, and Gc are shown in Fig. 6.

3.6. Age, Sex, and Age × Sex interaction

Given the differences in the magnitude and statistical significance of effects across age, we developed two

Table 8
Invariance model comparisons for 17 to 18 year olds

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	*p*	CFI	RMSEA	AIC
1. Configural	270.88	188				0.964	0.040	502.88
2. First-order factor loading	292.53	202	21.65	14	0.09	0.961	0.058	496.53
3. Subtest intercept	307.22	213	14.69	11	0.20	0.960	0.057	489.22
4. Subtest residual VAR/COV	317.12	231	9.90	18	0.94	0.963	0.052	463.12
5. Second-order loading	321.45	235	4.33	4	0.36	0.963	0.052	459.45
6. First-order unique VAR	325.42	240	3.97	5	0.55	0.963	0.051	453.42
7. *g* VAR	327.50	241	2.08	1	0.15	0.963	0.051	453.50
8. *g* mean diff (i.e. no diff in first-order means)	351.49	245	23.99	4	<0.01	0.954	0.057	469.49
9. *g*, Gv mean diff	338.31	244	13.18	1	<0.01	0.960	0.054	458.31
10. *g*, Gc, Gv mean diff	335.37	243	2.94	1	0.09	0.960	0.052	457.37
11. Gc, Gv mean diff	340.11	245	4.74	2	0.09	0.959	0.054	458.11
12. MIMIC model	190.14	107				0.965	0.053	316.14
13. MIMIC with sex effect on *g* removed	193.32	108	3.18	1	0.07	0.964	0.054	317.32

Note. Compare models 2 to 11 with the previous model in the table. Compare model 13 with model 12.

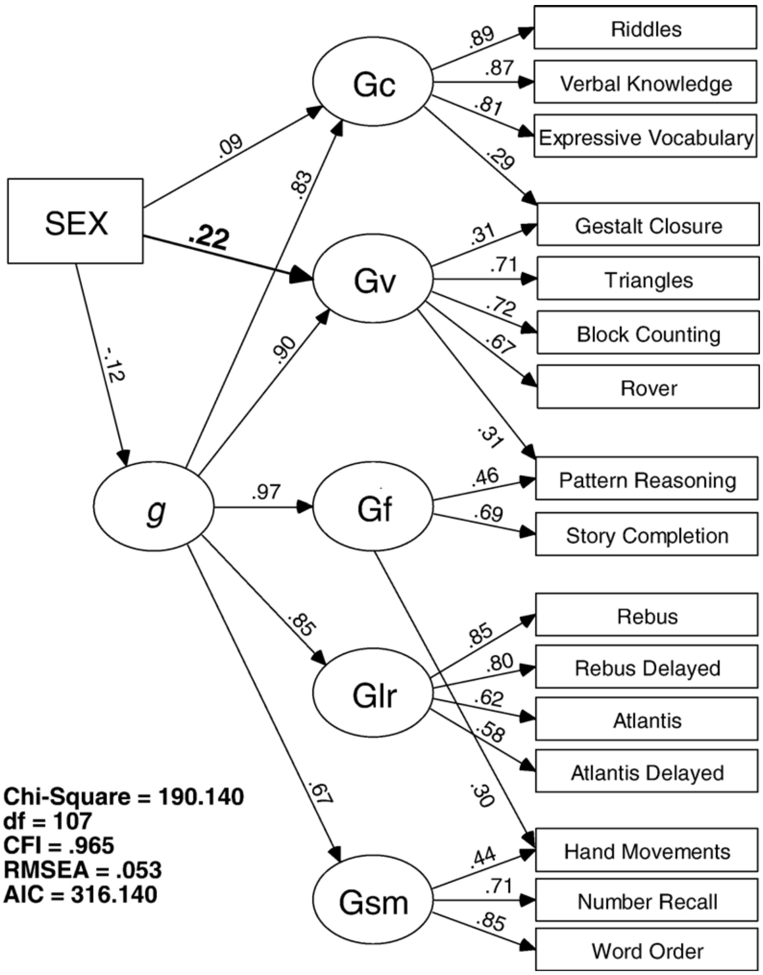

Fig. 6. The 17 to 18 year-old MIMIC model with statistically significant standardized effects of sex in bold and larger font. Negative paths from sex indicate a female advantage; positive paths from sex indicate a male advantage.

MIMIC models to test for a Sex × Age interaction. The first model tested the main effects of Age and Sex, and an interaction effect of Age × Sex (created by multiplying the sex variable times the centered age variable). For this model we used all of the data across the 6 to 18 year-old age range. Because previous research using these same KABC-II data suggested that the test is invariant in what it measures across the ages (Reynolds et al., in press) and because we found evidence for invariance across sex in all age groupings, we considered that the assumptions were met for the MIMIC model. An overall factor model across the ages was estimated with Age, Sex, and Age × Sex included as predictors. Direct paths from these variables to *g*, Gc, Gv, Glr and Gsm were included. Paths to Gf were not included for model identification purposes, but we did include a path from the variables to the Rover subtest. The model fit well: χ^2 (N=2375, df=127)=502.62; CFI=.984; RMSEA=.035; and AIC=666.62. The Gf unique variance was positive, but not statistically significant. Interestingly, the only statistically significant paths were from Sex to *g*, Gv, Gc, and Rover. None of the paths from Age or from the Age × Sex interaction variable was statistically significant. These findings were somewhat inconsistent with our findings from the MIMIC and MG-MACS models performed when the sample was divided into age groups because those findings suggested different statistically significant Sex effects at some age levels. The interaction analysis across all age groups suggests that there is no Sex × Age interaction, only significant main effects.

Next, we kept the Sex, Age, and Sex × Age variables in the model, but deleted all non-significant paths. Again the model was plausible: χ^2 (N=2375, df=139)= 516.12; CFI=.984; RMSEA=.034; and AIC=654.12. A $\Delta\chi^2$ of 13.5 with 12 df indicated that the model did not

Table 9
Status of substantive hypotheses in cognitive ability differences for boys and girls on the KABC-II, by age group

Age group	Summary of statistically significant differences			
	g variance	Broad ability variance	*g* means	Broad ability means
6–8	No	No	Girls higher	Boys higher on Gv, Gc
9–11	No	No	No	Boys higher on Gv, Gc
12–14	No	Girls more on Gsm	No	Boys higher on Gv, Gc
15–16	No	No	Girls higher	Boys higher on Gv, Gc
17–18	No	No	No	Boys higher on Gv

fit significantly worse than did the previous model. From this model we obtained the path estimates from Sex to *g* (β=−.09, b=−.40, SE=.11), Gv (β=.20, b=.81, SE=.09) Gc (β=.12, b=.66, SE=.10) and Rover (β=.07, b=.41, SE=.11). Girls showed an advantage on *g*; and when controlling for *g*, boys showed an advantage on Gv and Gc. Moreover, when controlling for *g* and Gv, boys showed an advantage on Rover. Rover appears to require some quantitative reasoning ability, and perhaps a separate quantitative reasoning factor would account for this difference.

3.7. Summary

Table 9 shows a summary of answers to the substantive questions to this research when the overall sample was divided into separate age groups. As shown in the table, girls varied more on one broad ability, Gsm, in one age group, 12 to 14. No other factor variance difference was statistically significant, suggesting generally homogenous construct variance across groups across ages. The one consistent statistically significant finding across every age group was that boys showed an advantage on Gv after controlling for the effects of *g*. At most ages, except for the 17 to 18 year age group, the boys also showed an advantage in Gc after controlling for the level of *g*. Last, at ages 6 to 8, and 15 to 16, girls showed a statistically significant advantage on *g*.

Because readers may be more familiar with standard IQ scales, we show these differences in standard IQ metric (i.e., M=100, SD=15) in Fig. 7. Although not all values were statistically significant, the graph shows considerable consistency in both the direction and magnitude of the sex differences. Moreover, our follow-up test for a Sex×Age interaction was not significant, further suggesting consistent Sex main effects across ages.

3.8. Discussion

Questions about possible differences in the level and variability in the cognitive abilities of males and females have been of interest to researchers for as long as human cognitive abilities have been measured. In the present study, we tested whether boys and girls differed in mean levels of *g* and broad cognitive abilities, and whether they differed in variance of those abilities. MG-MACS and MIMIC analyses of higher-order models of intelligence, consistent with contemporary intelligence theory, allowed for tests of whether differences in general intelligence, broad cognitive abilities, or both, accounted for mean differences in observed scores. Data

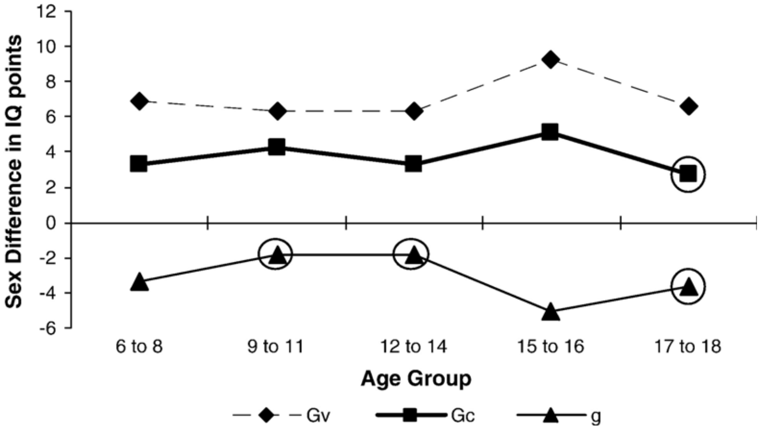

Fig. 7. Mean sex differences across ages for *g*, Gv, and Gc. Positive values indicate an advantage for boys, negative values show an advantage for girls. Circled values are not statistically significant. All values have been converted to a standard IQ score metric.

from a well-supported individually administered test of intelligence, the KABC-II, were used. The KABC-II is designed to assess general intelligence and five broad abilities consistent with three-stratum theory (Kaufman & Kaufman, 2004). Our findings suggested no statistically significant sex difference in variability in *g* or in broad cognitive abilities, except for Gsm at ages 12 to 14. In contrast, we found consistent statistically significant boy advantages in Gv means when controlling for *g*, fairly consistent statistically significant boy advantages in Gc means when controlling for *g*, and inconsistent statistically significant girl advantages in *g* means. When specifying a MIMIC model across the ages, only statistically significant main effects for sex were found on Gv, Gc, and *g*. No significant Sex × Age interaction was found. The estimates of mean differences were the same for MG-MACS or MIMIC models when the factor loadings and variances were invariant across the groups.

3.9. Variability of latent cognitive abilities

Previous research suggests that boys are more variable than girls in *g* and IQ scores (e.g., Arden & Plomin, 2006; Deary et al., 2003). Our findings are inconsistent with such previous research. Except for higher variability in Gsm for girls at ages 12 to 14, the variability in broad abilities and in *g* was not statistically different across sex at any age. These findings suggest that boys and girls rely on a similar range of abilities when presented with cognitive tasks. Differences in our findings from previous studies (e.g., Arden & Plomin, 2006; Deary et al., 2003) may be a result of different sample sizes; the other two studies used extremely large sample sizes (*n*s 7748 to 87,498), which would increase the likelihood of achieving statistical significance. Our findings may also be the result of the use of a test standardization sample in our research. Although the KABC-II included children and adolescents from both ends of the intelligence continuum in its standardization, as well as those with disabilities, relatively few appear in a sample of 2375. Thus, our data may not be suited to a strong test of the hypothesis that males are overrepresented on the two ends of the cognitive spectrum (and thus show greater variance in intelligence).

3.10. Mean differences in latent cognitive abilities

Our findings were consistent with previous research in suggesting that the boy advantage on tests of visual–spatial ability is the result of true differences in visual–spatial ability. These differences cannot be accounted for by differences in *g*, and were demonstrated at every age level in the KABC-II data. This advantage also appeared to be related to the broader visual–spatial construct, and not to more specific abilities (except for at ages 9 to 11 where higher scores for boys on the Rover test were attributable to Gv, a more specific factor not accounted for by our model, or both).

Our general findings were consistent with previous research that showed an advantage for boys in Gc. Camarata and Woodcock (2006) found that 5 to 18 year-old boys showed an advantage in Gc. Although in our study there was not a statistically significant advantage at ages 17 and 18 in the MG-MACS models, our findings from the MIMIC models are consistent with those of Camarata and Woodcock in that there was no significant Age × Sex interaction effect, but a significant main effect of Sex.

Boys in our study did not show an advantage in *g* at any age group. If indeed males are considered to have a higher general intelligence, then perhaps the developmental theory proposed by Lynn (1994, 1999) could explain the girl advantage at younger ages. Specifically, according to Lynn's theory, girls may show advantages, or similar levels, of *g* at younger ages, but as boys' maturity accelerates and catches up to girls', the trend should begin to switch to a boy advantage around ages 15 to 16. Therefore, in our study, according to Lynn's theory, boys should have started to even out in *g* levels by ages 15 and 16 with a more visible advantage evident at ages 17 and 18. For example, Colom and Lynn (2004) found that although girls performed higher than boys on a global IQ score at the ages of 12 to 13, boys started to show an upward trend compared to girls after the age of 15, surpassing girls at age 16, and eventually showed an advantage of 4.3 IQ points at age 18. In our study, however, girls showed a statistically significant advantage in *g* at ages 15 and 16, and an advantage at ages 17 and 18 (albeit, statistically non-significant), so the upward trend observed in the Colom and Lynn study was not observed in the present study. At ages 17 to 18, girls showed an advantage that was approximately equal to 3.6 IQ points. Interestingly, boys also showed a consistent advantage in Gv (~6.6 IQ points at age 18) and Gc (~2.7 IQ points at age 18) throughout childhood when these abilities were tested separately from *g*. Most important to note, however, is that the differences between boys and girls were consistently in the same direction and of similar magnitude from the ages of 6 to 18. Here, a sex by age interaction effect was not supported, a finding inconsistent with Lynn's developmental theory.

One explanation for the inconsistencies in our findings and some of the previous literature is the use of MG-MACS and MIMIC models. The use of latent variable models is important because the models are consistent with theories of cognitive abilities. Furthermore, MG-MACS analyses allow for tests of measurement equivalence *before* tests of group differences are performed. Lynn's (1998) conclusion that males show an IQ advantage whether general intelligence is defined by the sum of group factors, global IQ score, or first principal component may be correct. To our knowledge males have yet to show an advantage in a second-order *g* factor that has strictly satisfied the necessary condition of factorial invariance, only after which group comparisons are considered valid. It is necessary, however, that our findings be replicated in other intelligence batteries and in older age groups.

Another possible difference in our findings and findings from previous research may be attributed to the measures used to assess intelligence. The incorporation of theory in the development of the KABC-II represents a new trend in intelligence test development (Kamphaus, Windsor, Rowe, & Kim, 2005). The application of theory is important to the development of intellectual measures because it provides clarity in development and allows test developers to design tests specifically aimed at tapping into particular abilities. Thus, the KABC-II represents an assessment battery that was grounded in three-stratum theory and tests that were specifically designed to measure aspects of five different broad abilities (Kaufman, Kaufman, Kaufman-Singer, & Kaufman, 2005). For example, in the past, many of the most popular intelligence tests have been atheoretical. A lack of theoretical basis may well have resulted in ill-defined constructs and narrow or unrepresentative composites. Research using such tests may have therefore tested incomplete representations of the intended broad and general abilities. Research using latent variable methods with theoretically derived instruments may allow for more confident conclusions about sex differences in cognitive abilities.

Lastly, an alternative explanation was brought to our attention concerning our findings of *g* differences favoring girls. One possibility is that the girl advantage could have resulted from a systematic elimination of test items favoring boys in the development of the KABC-II (cf. Jensen, 1998). In the present research, all of the available evidence suggests that this possibility is not correct. It is certainly the case that "analyses of differential item functioning were carried out to identify any items that were *differentially* difficult by sex, ethnic group, or SES" (Kaufman & Kaufman, 2004, p. 84, emphasis added). DIF analyses do not remove items simply because groups show mean differences; instead, items are removed if they are differentially difficult for one group or the other, *after controlling for differences in the latent trait being measured* (Embretson & Reise, 2000, pg. 252). Item bias on the KABC-II was dealt with comparably for boys and girls (M. H. Daniel, personal communication, March 3, 2007). All available evidence suggests that our findings are not a result of the test development process, but instead are a result of true sex differences in latent constructs. Nevertheless, it is possible that methods used to develop and to eliminate items could have some unknown effect on findings of sex differences in general and broad abilities. Of course, such a caveat applies to all such research, using any test.

3.11. Limitations

This study was limited in that the data were cross-sectional. The conclusions would have been stronger with longitudinal data so that the dynamic development of human cognitive abilities could have been modeled across sex. Consideration of age as a variable served as the best "stand-in" available for understanding developmental differences. Our attempt to control for developmental differences was made by splitting the sample into different age groups as well as by testing a Sex × Age interaction. It is important to note that although the data were not longitudinal, the KABC-II sample was representative of the United States population at all ages and thus the participants were similar to U.S. children in numerous characteristics across the ages.

This study was also limited in that we were constrained by the subtests included in the KABC-II. The psychometric properties and internal validity of the KABC-II are excellent and it is well-aligned with three-stratum theory (Carroll, 1993, 1997); however, the KABC-II only provides measures of five different broad abilities with only one pure measure of fluid reasoning.

As already noted, although the standardization sample used in this research included a representative numbers of individuals with disabilities (e.g., mental retardation) the sample was too small to include large numbers of individuals at the extremes of the IQ continuum. Thus our data could not have allowed for as strong a test of sex differences in variances as research focused on larger samples or populations (e.g., Deary et al., 2003).

3.11.1. Alternative models

A final limitation of this research is that we focused on sex differences based solely on the three-stratum model of intelligence. Although this was the model used to develop the KABC-II, it is certainly possible that models representing other theories of intelligence would show different findings for sex differences. To test this possibility, we analyzed several additional MIMIC models with alternative factor structures. First, sex differences were investigated using a bi-factor (Holzinger & Swineford, 1937) or nested factors model (Gustaffson & Balke, 1993). In the bi-factor model, the general factor is referred to as *G* because it occupies the same factor space as the first-order factors, and all subtests load directly onto *G*. The Gf factor was collapsed in this model. Pattern Reasoning loaded on a first-order Gv factor and *G*, and Story Completion was not specified to load on a first-order factor and loaded only on *G* (it was allowed to load on the Gc factor, but the loading was negligible). Five first-order factors were specified: one *G* factor and four broad ability factors, Gv, Gc, Glr, and Gsm. A Sex variable was specified to have direct paths to all of the broad ability factors and *G* factor simultaneously, and the entire sample was used.

The model fit well: χ^2 (N=2375, df=97)=359.31; CFI=.985; RMSEA=.034; and AIC=505.31. Girls demonstrated a statistically significantly advantage on *G* (β=−.11, b=−.50, SE=.12), whereas boys showed a statistically significant advantage on Gv (β=.40, b=.64, SE=.13) and Gc (β=.23, b=.74, SE=.11). The sex differences for Gsm and Glr were not statistically significant.

For four of the five age and sex levels used in this research, the path from *g* to Gf was close to 1.0, resulting in a nonsignificant unique variance for the Gf factor. Although *g* and Gf factors are often indistinguishable, we wanted to make sure that the results would be similar if the Gf factor variance was fixed to zero or if the factor was collapsed in the second-order model. Before these two models were specified, for purposes of comparison, a model with paths from Sex to Gc, Gv, Gsm, and Glr, but setting the path from Sex to Gf to zero, was estimated. Again, the entire sample was used in these analyses. The model fit well [χ^2 (N=2375, df=104)=394.92; CFI=.983; RMSEA=.034; AIC=526.92] and the Gf variance was not statistically significantly different from zero. Girls showed an advantage in *g* and boys showed advantages in Gv and Gc; the Sex to Gsm and Glr paths were both nonsignificant. Next, another model with the Gf variance constrained to zero was estimated. The model fit was not worse than the previous model ($\Delta\chi^2$=1.23, Δdf=1, p=.27, AIC=526.15), and all paths were consistent with the previous model. We also estimated a model in which the Gf factor was deleted from the model. Pattern Reasoning loaded on *g* and Gv, Story Completion loaded directly on *g*, and Hand Movements loaded on *g* and on Gsm. As expected, there were no differences in the model fit or other findings. The path from Sex to *g* was statistically significant in favor of girls, the path from Sex to Gv and Gc was significant in favor of boys, and the Sex to Gsm and Glr paths were not statistically significant.

It is tempting to think that the differences for boys and girls in *g* are due primarily to differences in Gsm and Glr because additional broad ability differences were allowed for the Gv and Gc factors. Another set of analyses was performed in which all Gsm and Glr subtests were removed. Sex differences were allowed for *g*, Gc and Gv, but the path to Gf from Sex was set to zero. Girls demonstrated a statistically significant advantage in *g* (β=−.11, b=−.47 SE=.13), boys showed a statistically significant advantage on Gv (β=.22; b=.87, SE=.11) and Gc (β=.13; b=.72, SE=.13).

A model with three broad abilities—Verbal, Performance/Perceptual, and Memory (plus *g*)—was also tested. The Gv and Gf factors were combined. Two Glr tests (Atlantis and Atlantis Delayed) were placed on the Verbal factor, and two (Rebus and Rebus Delayed) were placed on the Performance/Perceptual factor. The model fit well [χ^2 (N=2375, df=109)=846.12; CFI=.962; RMSEA=.047; AIC=968.12], although less well than any preceding model. The results were consistent with the primary findings of this research: boys showed a statistically significant advantage on the Performance/Perceptual (β=.05; b=.20, SE=.09) and Verbal (β=.11; b=.56, SE=.13) factors, whereas girls showed a statistically significant advantage on *g* (β=−.10; b=−.44, SE=.14). Similar results were shown when the Glr subtests were deleted from the analysis.

Finally, the KABC-II can be interpreted from two theoretical orientations: three-stratum theory or Luria's neuropsychological model; we estimated a model more consistent with the Luria interpretation. In the Luria interpretation of the KABC-II, Gv equals Simultaneous Processing, Gf equals Planning, Gsm equals Sequential Processing, and Glr equals Learning. The global IQ score without the Gc/Knowledge subtests scores is referred to as a Mental Processing Index in the KABC-II. Although Gc tests are listed as Knowledge in the Luria theoretical interpretation of the test, to be more consistent with the Luria model, the Gc subtests were deleted from our model. The *g* factor was included in the model to effectively partial it out from the first-order factors. Paths from Sex to *g* (or General Mental Processing) Sequential, Simultaneous, and Learning were included; however, in order to properly identify the model, one path was fixed to zero. We alternated these constraints across a series of models allowing for different paths to the first-order factors to be constrained in different models. The end result was a statistically significant girl advantage in General Mental Processing (*g*; $\beta=-.10$, $b=-.43$ SE=.13), and a boy advantage in Simultaneous processing (Gv; $\beta=.25$, $b=.98$, SE=.10).

It is reassuring that the results of all of these supplemental analyses were so consistent with the primary MG-MACS and MIMIC analyses. We consider this consistency a result of the use of a latent variable, as opposed to emergent variable, approach. Latent variables remove *g*, unique and error variance from consideration of sex differences in broad abilities, and also remove the unique aspects of the broad abilities from consideration of sex differences in *g*. As a result, the findings of research using latent variables should be less dependent on idiosyncrasies in samples, factor structures, or subtests. Our demonstration of factorial invariance before testing for differences also lends confidence to the comparisons of the *factors* from the model used in this research. Nevertheless, there are many other possible factor structures that were not evaluated in this research (e.g., extended theory of fluid and crystallized cognitive abilities (Horn & Blankson, 2004); Vernon's v:ed and k:m model (1965), or the verbal–perceptual–image rotation model (Johnson & Bouchard, 2005); such models may have produced a different pattern of sex differences than those shown here.

3.12. Directions for future research

Previous research has suggested that sex differences in cognitive abilities may be the result of educational attainment (Dolan et al., 2006). The sample used in our study included children and adolescents ages 6 through 18 who were all participants in mandatory education programs in the United States. Therefore, educational attainment is unlikely to account for the mean differences found in ability. This conclusion may be more debatable, however, at ages 17 to 18, when students begin to have more choice in their educational programs, with girls and boys perhaps selecting different coursework. Future research should continue to investigate these differences, but should also attempt to account for other individual differences, such as SES, ethnic backgrounds, and high school program of study.

Future research needs to continue to model latent abilities as such. Analyses that model cognitive abilities as linear composites of test scores are inconsistent with intelligence theory. Intelligence theory suggests that latent dimensions underlie performance on cognitive tests. Good science requires that these latent dimensions be structured so that they relate to the observed scores in the same way in different groups before meaningful group comparisons are made. Moreover, observed scores are fraught with construct irrelevant variance that may well play a role in findings of mean differences on observed measures, and may be a reason for inconsistencies in past research findings on group differences in intelligence.

Last, when the assumptions for the MIMIC model were met, both MG-MACS and MIMIC analyses resulted in identical estimates of mean differences, a consistency expected, and confirmed. These findings show that MIMIC models may be used in future sex differences research if the assumptions of those models are met. Regardless of which type of model is used, future research should continue to focus on latent constructs if the constructs are hypothesized as being latent. At the same time, it is important to try to elucidate the variants and different assumptions of various latent variable models.

3.13. Summary

The Kaufman Assessment Battery for Children—Second Edition was used to investigate sex differences in latent cognitive abilities. The standardization sample was a nationally representative (U.S.) sample of children and adolescents, ages 6 to 18. MG-MACS and MIMIC models of the KABC-II factor structure, a test aligned with three-stratum theory, demonstrated that at all ages boys showed a mean advantage in a latent visual–spatial ability (Gv) compared to girls when controlling for the effects of *g*. Moreover, boys showed a mean advantage in a latent crystallized ability (Gc) at ages 6 to 16, and a statistically non-significant advantage at 17 to 18 when controlling for *g*. Girls demonstrated an advantage on the latent second-order *g* factor, although this difference was statistically significant only at ages 6 to 7 and 15 to 16. Nevertheless, a test of Age × Sex interaction was not significant, suggesting only significant main effects of sex on Gv, Gc, and *g*. Last, other than Gsm being more variable for girls at ages 12 to 14, our findings suggest that boys and girls rely on a similar range of latent abilities when performing cognitive tasks. These findings suggest that sex differences in cognitive abilities are present and consistent in direction and magnitude for children ages 6 to 18.

Appendix A

These tables show the means and standard deviations, along with information about subtest mean differences and estimated effects sizes, for the KABC-II subtests for boys and girls in each age group. Results of MANOVAs and Box's tests of covariance matrix equality are presented below each table.

Table A1
Sample sizes, subtest means, standard deviations, effect sizes, MANOVA, and Box's test results for 6 to 8 year olds

Subtest	Boys			Girls				
	M	SD	*N*	M	SD	*N*	*t*	*d*
Atlantis Delayed	9.79	2.81	295	9.85	2.91	293	−0.28	−0.02
Atlantis	9.95	3.15	304	10.25	3.27	296	−1.14	−0.09
Block Counting	10.28	3.13	304	9.77	2.95	296	2.03	0.17
Expressive Vocabulary	9.88	3.02	304	9.82	2.97	296	0.25	0.02
Gestalt Closure	10.42	2.65	230	9.72	2.94	229	2.70*	0.25
Hand Movements	9.98	2.93	304	10.29	2.68	296	−1.35	−0.11
Number Recall	10.10	2.78	304	10.46	2.93	296	−1.55	−0.13
Pattern Reasoning, Untimed	9.58	3.06	304	9.91	2.85	296	−1.38	−0.11
Rebus Learning, Delayed	9.88	3.03	288	10.18	2.92	288	−1.19	−0.10
Riddles	10.07	2.92	304	9.92	3.04	296	0.59	0.05
Rebus Learning	9.94	3.28	304	10.39	3.05	296	−1.74	−0.14
Rover	10.33	2.99	304	9.90	2.75	296	1.84	0.15
Story Completion, Untimed	9.94	2.86	304	10.29	2.86	296	−1.50	−0.12
Triangles, Untimed	10.28	3.01	304	9.84	2.90	296	1.79	0.15
Verbal Knowledge	10.01	2.84	304	9.97	3.03	296	0.17	0.01
Word Order	9.66	2.87	304	10.13	2.88	296	−1.99	−0.16

*$p<.01$.
Note. Above positive signs indicate higher scores for boys, negative signs indicate higher scores for girls.
MANOVA results indicated that sex did not significantly affect the combination of subtests $F(16, 420)=1.63$, $p=.06$.
Box's test revealed that equality of covariance matrices could be assumed: $F(136, 581{,}354)=1.07$, $p=.26$.

Table A2
Sample sizes, subtest means, standard deviations, effect sizes, MANOVA, and Box's test results for 9 to 11 year olds

Subtest	Boys			Girls				
	M	SD	*N*	M	SD	*N*	*t*	*d*
Atlantis Delayed	9.85	2.89	300	9.83	2.95	297	0.10	0.01
Atlantis	9.97	3.10	301	9.68	2.94	299	1.15	0.09
Block Counting	9.81	2.94	301	9.78	3.03	299	0.16	0.01
Expressive Vocabulary	9.95	3.17	301	9.55	2.75	299	1.69	0.14
Gestalt Closure	10.50	2.65	200	9.82	3.12	215	2.38	0.23
Hand Movements	9.97	2.87	301	10.11	3.00	299	−0.59	−0.05
Number Recall	10.24	2.79	301	10.14	2.96	299	0.41	0.03
Pattern Reasoning, Untimed	9.90	3.16	301	9.68	2.89	299	0.91	0.07
Rebus Learning, Delayed	9.66	2.94	287	10.22	2.76	288	−2.33	−0.19
Riddles	10.20	3.16	301	9.78	3.00	299	1.70	0.14
Rebus Learning	9.87	3.09	301	10.26	2.82	299	−1.63	−0.13
Rover	10.56	2.86	301	9.58	2.96	299	4.11*	0.34
Story Completion, Untimed	9.72	2.82	301	9.96	2.92	299	−1.02	−0.08
Triangles, Untimed	10.50	2.83	301	9.73	3.04	299	3.22*	0.26
Verbal Knowledge	10.20	3.21	301	9.76	2.80	299	1.79	0.15
Word Order	9.72	3.00	301	10.09	2.79	299	−1.58	−0.13

*$p<.01$.
Note. Above positive signs indicate higher scores for boys, negative signs indicate higher scores for girls.
MANOVA results indicated that sex did significantly affect the combination of subtests $F(16, 383)=2.35$, $p=.002$.
Box's test revealed equality of covariance matrices could be assumed: $F(136, 482{,}878)=1.09$, $p=.22$.

Table A3
Sample sizes, subtest means, standard deviations, effect sizes, MANOVA, and Box's test results for 12 to 14 year olds

Subtest	Boys			Girls				
	M	SD	*N*	M	SD	*N*	*t*	*d*
Atlantis Delayed	10.02	2.78	296	9.86	2.81	297	0.72	0.06
Atlantis	10.21	2.97	302	9.87	3.03	298	1.39	0.11
Block Counting	10.01	3.21	302	9.65	2.90	298	1.46	0.12
Expressive Vocabulary	9.98	3.09	302	9.49	2.90	298	2.01	0.16
Gestalt Closure	10.18	2.97	218	10.00	2.98	215	0.66	0.06
Hand Movements	9.85	2.86	302	9.98	2.91	298	−0.58	−0.05
Number Recall	10.09	2.68	302	10.19	3.04	298	−0.45	−0.04
Pattern Reasoning, Untimed	9.79	3.17	302	9.74	2.82	298	0.18	0.01
Rebus Learning, Delayed	9.99	3.09	300	10.11	3.07	296	−0.47	−0.04
Riddles	10.17	3.19	302	9.92	3.10	298	0.96	0.08
Rebus Learning	9.82	3.13	302	10.11	3.12	298	−1.16	−0.09
Rover	10.57	3.17	302	9.42	2.77	298	4.72*	0.39
Story Completion, Untimed	9.68	2.85	302	10.18	2.91	298	−2.15	−0.18
Triangles, Untimed	10.32	2.89	302	9.76	2.90	298	2.38	0.19
Verbal Knowledge	10.15	3.10	302	9.68	3.08	298	1.83	0.15
Word Order	9.81	2.80	302	9.69	2.82	298	0.52	0.04

*$p < .01$.
Note. Above positive signs indicate higher scores for boys, negative signs indicate higher scores for girls.
MANOVA results indicated that sex did significantly affect the combination of subtests $F(16, 409) = 2.84$, $p < .001$.
Box's test revealed equality of covariance matrices could be assumed: $F(136, 555{,}165) = 1.26$, $p = .02$.

Table A4
Sample sizes, subtest means, standard deviations, effect sizes, MANOVA, and Box's test results for 15 to 16 year olds

Subtest	Boys			Girls				
	M	SD	*N*	M	SD	*N*	*t*	*d*
Atlantis Delayed	9.96	2.80	146	10.20	3.02	149	−0.71	−0.08
Atlantis	10.04	3.36	146	10.10	2.88	154	−0.17	−0.02
Block Counting	10.33	3.09	146	9.59	2.69	154	2.21	0.26
Expressive Vocabulary	9.99	3.05	146	9.66	2.77	154	0.98	0.11
Gestalt Closure	10.13	2.95	99	9.62	3.14	114	1.21	0.17
Hand Movements	9.55	3.02	146	10.28	2.89	154	−2.14	−0.25
Number Recall	9.67	2.86	146	10.08	2.97	154	−1.21	−0.14
Pattern Reasoning, Untimed	10.06	3.04	146	10.26	2.88	154	−0.58	−0.07
Rebus Learning, Delayed	9.77	2.94	145	10.57	3.00	148	−2.31	−0.27
Riddles	10.38	3.36	146	10.31	3.14	154	0.21	0.02
Rebus Learning	9.77	2.74	146	10.64	2.88	154	−2.68*	−0.31
Rover	10.34	3.09	146	9.59	2.84	154	2.18	0.25
Story Completion, Untimed	9.32	2.78	146	9.88	2.67	154	−1.76	−0.20
Triangles, Untimed	10.34	2.59	146	9.85	2.62	154	1.64	0.19
Verbal Knowledge	10.13	3.02	146	9.99	3.03	154	0.41	0.05
Word Order	9.75	2.83	146	10.16	2.91	154	−1.24	−0.14

*$p < .01$.
Note. Above positive signs indicate higher scores for boys, negative signs indicate higher scores for girls.
MANOVA results indicated that sex did significantly affect the combination of subtests $F(16, 189) = 2.27$, $p < .001$.
Box's test revealed that equality of covariance matrices could be assumed: $F(136, 126, 069) = .97$, $p = .59$.

Table A5
Sample sizes, subtest means, standard deviations, effect sizes, MANOVA, and Box's test results for 17 to 18 year olds

Subtest	Boys			Girls				
	M	SD	*N*	M	SD	*N*	*t*	*d*
Atlantis Delayed	9.57	3.28	131	10.01	3.02	142	−1.16	−0.14
Atlantis	9.56	3.44	133	9.70	3.54	142	−0.32	−0.04

Table A5 (*continued*)

Subtest	Boys			Girls				
	M	SD	*N*	M	SD	*N*	*t*	*d*
Block Counting	10.16	3.32	133	9.92	2.78	142	0.64	0.08
Expressive Vocabulary	9.99	2.95	133	9.99	3.08	142	0.00	0.00
Gestalt Closure	10.24	3.11	102	9.90	3.12	116	0.80	0.11
Hand Movements	9.97	2.95	133	10.07	2.87	142	−0.29	−0.03
Number Recall	9.56	3.17	133	9.68	2.79	142	−0.33	−0.04
Pattern Reasoning, Untimed	10.02	3.37	133	10.26	2.82	142	−0.66	−0.08
Rebus Learning, Delayed	9.55	3.14	132	10.48	2.83	139	−2.58	−0.31
Riddles	10.02	2.97	133	10.23	2.95	142	−0.59	−0.07
Rebus Learning	9.52	3.25	133	10.73	2.91	142	−3.25*	−0.39
Rover	10.43	3.28	133	9.45	3.05	142	2.56	0.31
Story Completion, Untimed	9.78	2.71	133	10.02	2.95	142	−0.70	−0.08
Triangles, Untimed	9.70	2.81	133	9.38	2.43	142	1.01	0.12
Verbal Knowledge	9.98	3.12	133	10.00	3.01	142	−0.06	−0.01
Word Order	10.05	3.07	133	9.85	2.86	142	0.56	0.07

*$p < .01$.

Note. Above positive signs indicate higher scores for boys, negative signs indicate higher scores for girls.

MANOVA results indicated that sex did not significantly affect the combination of subtests $F(16, 195) = 2.02$, $p = .014$.

Box's test revealed that equality of covariance matrices could be assumed: $F(136, 131, 454) = 1.00$, $p = .50$.

References

Akaike, H. A. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332.

Arbuckle, J. L. (2003). *Amos 5.0 update to the Amos user's guide.* Chicago: Smallwaters.

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, *41*, 39–48.

Bartholomew, D. J. (2004). *Measuring intelligence: Facts and fallacies.* New York: Cambridge University Press.

Braden, J. P., & Ouzts, S. M. (2005). Review of the Kaufman Assessment Battery for Children, Second Edition. In R. A. Spies, & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 517–520). Lincoln, NE: Buros Institute of Mental Measurements.

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

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, *13*, 287–321.

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, *34*, 231–252.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 122–130). New York: Guilford.

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

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.

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

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

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

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, *62*, 61–84.

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

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

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 239–247.

Halpern, D. F. (1997). Sex differences in intelligence. *American Psychologist*, *52*, 1091–1102.

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

Hancock, G. E. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, *30*, 91–105.

Härnqvist, K. (1997). Gender and grade differences in latent ability variables. *Scandinavian Journal of Psychology*, *38*, 55–62.

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A re-analysis of Feingold's results. *Review of Educational Research*, *63*, 94–105.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–55.

Horn, J. L., & Blankson, N. (2004). Foundations for better understanding of cognitive abilities. In D. P. Flanagan, & P. L. Harrison

(Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 41−68)., 2nd ed. New York, NY: Guilford.

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

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.

Hulick, P. A. (1998). A structural factor analysis of gender and age differences in cognitive ability. In McArdle, & Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 247−262). Mahwah, NJ: Erlbaum.

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.

Jensen, A. R. (1998). *The g factor.* Westport, CT: Prager.

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It's verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393−416.

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

Kamphaus, R. W., Pierce Windsor, A., Rowe, E. W., & Kim, S. (2005). A history of intelligence test interpretation. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 23−38)., 2nd ed. New York, NY: Guilford.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children—Second Edition: Technical manual.* Circle Pines, MN: American Guidance Service.

Kaufman, J. C., Kaufman, A. S., Kaufman-Singer, J., & Kaufman, N. L. (2005). The Kaufman Assessment Battery for Children—Second Edition and the Kaufman Adolescent and Adult Intelligence Test. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 344−369)., 2nd ed. New York, NY: Guilford.

Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 581−614)., 2nd ed. New York, NY: Guilford.

Keith, T. Z. (2006). *Multiple regression and beyond.* Boston, MA: Allyn and Bacon.

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher-order, multi-sample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fourth edition: What does it measure? *School Psychology Review*, *35*, 108−127.

Kimura, D. (2004). Human sex differences in cognition, fact, not predicament. *Sexualities, Evolution, & Gender*, *6*, 45−53.

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

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

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

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

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

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 advanced progressive matrices in college students. *Personality and Individual Differences*, *37*, 219−223.

Mackintosh, N. J. (1996). Sex differences and IQ. *Journal of Biosocial Science*, *28*, 559−572.

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.

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136−181)., 2nd ed. New York, NY: Guilford.

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

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In A. Sayer, & L. Collins (Eds.), *New methods for the analysis of change* (pp. 203−240). Washington, DC: American Psychological Association.

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 intelligence: Tribute to Arthur R. Jensen* (pp. 187−222). Oxford: Elsevier Science.

Reynolds, M. R., Keith, T. Z., Fine, J. G., Fisher, M. F., & Low, J. (in press). Confirmatory factor structure of the Kaufman Assessment Battery for Children—Second Edition: Consistency with Cattell–Horn–Carroll theory. *School Psychology Quarterly, 22*.

Rosén, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction*, *5*, 37−62.

Steenkamp, J. -B., E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78−90.

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling*, *5*, 411−419.

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119−170). Greenwich, CT: Information Age.

van der Sluis, S., Derom, C., Thiery, E., Bartels, M., Polderman, T., Verhulst, F. C., et al. (in press). Sex differences on the WISC-R in the Belgium and the Netherlands. *Intelligence*. doi:10.1016/j.intell.2007.01.003

van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J., Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence*, *34*, 273−289.

Vernon, P. E. (1965). Ability factors and environmental influences. *American Psychologist*, *20*, 723−733.

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

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