Intelligence 48 (2015) 96–108

Contents lists available at ScienceDirect

Intelligence

Gender differences in latent cognitive abilities in children aged 2 to 7

Mohammed H. Palejwala *, Jodene Goldenring Fine

Michigan State University, USA

ARTICLE INFO

Article history:
Received 16 April 2014
Received in revised form 20 September 2014
Accepted 3 November 2014
Available online 26 November 2014

Keywords:
Gender differences
General intelligence
Broad cognitive abilities
MG-CFA models
Developmental differences

ABSTRACT

Gender differences in the latent cognitive abilities underlying the Wechsler Primary and Preschool Scale of Intelligence—Fourth Edition (WPPSI-IV) were investigated in children aged 2 to 7. Multiple-group confirmatory factor analysis was used to verify the measurement invariance of the WPPSI-IV factor model in boys and girls. Then the magnitude of gender differences in the means and variances of the abilities was estimated. Multiple-indicator multiple-cause models were implemented to explore whether the magnitude of these differences varied across age. Girls aged 2 to 7 demonstrated higher general intelligence. Girls aged 4 to 7 demonstrated an advantage in processing speed. A gender difference favoring boys in visual processing was absent in ages 2 to 3 but emerged in ages 4 to 7. Gender differences in fluid reasoning, short-term memory, and comprehension-knowledge were not found. The variability of any of the abilities did not differ among girls and boys. These results indicate that gender differences in cognitive abilities emerge in early childhood, which may contribute to gender differences in later educational outcomes.

1. Introduction

The research on gender differences in cognitive abilities is marked by inconsistency. Even for those gender differences receiving consistent support in adults (e.g. a male advantage in visual–spatial ability), the age at which these differences emerge in childhood is unclear. The purpose of the present study is to investigate gender differences in cognitive abilities in children aged 2 to 7 years with the goal of determining when these differences appear. The instrument used to examine gender differences is the fourth edition of the Wechsler Primary and Preschool Scale of Intelligence (WPPSI-IV), one of the most widely used measures of intelligence for young children (Raiford & Coalson, 2014). The following section discusses the factors that explain inconsistencies in the literature on gender differences in cognitive abilities. Methodological differences among studies along with population heterogeneity may contribute to discrepancies in the extant literature.

1.1. Inconsistencies in previous research

Researchers have historically examined gender differences by comparing male and female scores on single tests or composites of multiple tests. These types of scores are referred to as observed scores. Observed scores contain measurement error and unique variance. In contrast, latent variables are estimates of cognitive abilities using structural equation modeling that remove these sources of unreliability and invalidity. Latent variables are less influenced by the mix of tests used to estimate them and are considered to be purer measures of the construct of interest. Studies using both observed and latent variable methods to examine gender differences in the same data set have shown that these methods produce different conclusions (Härnqvist, 1997; Steinmayr, Beauducel, & Spinath, 2010), supporting the need to use a latent variable approach to investigate gender differences in cognitive abilities.

* Corresponding author at: Michigan State University, 435 Erickson Hall, East Lansing, MI 48824-1034, USA. Tel.: +1 517 432 1683.
E-mail addresses: palejwal@msu.edu (M.H. Palejwala), finej@msu.edu (J.G. Fine).

http://dx.doi.org/10.1016/j.intell.2014.11.004
0160-2896/

Another advantage to using a latent variable methodology is that one can investigate the assumption that a test measures constructs in the same way across groups. This assumption is called measurement invariance and is a prerequisite to comparing scores reflecting the constructs. Studies examining the measurement invariance of cognitive ability tests across gender have sometimes found that the instruments only partially meet criteria for measurement invariance (Immekus & Maller, 2010; Keith, Reynolds, Roberts, Winter, & Austin, 2011). Therefore, measurement invariance of a cognitive test battery across gender should not be assumed and needs to be examined before comparing male and female scores on the battery.

Cognitive tests may not only measure a construct differently between groups, but additionally they may not always measure the ability that they intend to measure. Discrepancies in the literature on gender differences in cognitive abilities may arise from discrepancies in how cognitive abilities are operationalized. One frequently used theory for operationalizing the cognitive abilities that intelligence tests measure is Cattell–Horn–Carroll (CHC) theory (Keith & Reynolds, 2010). CHC theory is a taxonomy of cognitive abilities based on factor analysis of more than 460 data sets and is arguably among the best supported taxonomies of cognitive abilities (McGrew, 2009).

CHC theory defines cognitive abilities at three levels, or strata, of generality. The lowest level describes cognitive abilities with the most specificity and consists of more than 50 abilities called narrow abilities (stratum I). The narrow abilities can be classified into at least 7 abilities, which are called broad abilities (stratum II). The highest level describes cognitive abilities at the most general level and consists of one ability: general intelligence, or g (stratum III). The structure of CHC theory can be described by a second-order factor model, in which the broad abilities account for covariation among the narrow abilities, and g accounts for covariation in the broad abilities. Because the structure of the current version of the instrument used in this study is based on CHC theory, and it is a well-supported theory, it is used to define the cognitive abilities measured in this study.

Another methodological difference that may explain discrepancies in the gender differences literature is whether or not researchers account for g when comparing males and females on specific abilities. If g is not accounted for, gender differences in specific abilities may in reality reflect differences in general cognitive development. For this reason, g is controlled in the current study. Studies have found that the magnitude of gender differences in specific abilities can vary before and after controlling g, underlining the need to account for g in this type of investigation (Burns & Reynolds, 1988; Kaiser & Reynolds, 1985).

A non-methodological difference that likely contributes to discrepancies in the literature is population heterogeneity. Specifically, gender differences in cognitive abilities vary by age. Cross-sectional studies have found that gender differences in cognitive abilities measured by the same instrument emerge and diminish across the lifespan (Keith, Reynolds, Patel, & Ridley, 2008; Keith et al., 2011; Reynolds, Keith, Ridley, & Patel, 2008). These studies used instruments that demonstrate measurement invariance across ages, so the change in gender differences in cognitive abilities cannot be attributed to a change in the way the abilities are measured.

Based on this overview of the factors that contribute to discrepancies in the literature on gender differences in cognitive abilities, the strongest studies: (a) verify that their instrument measures cognitive abilities in the same way across gender, (b) estimate abilities at the latent variable level, (c) use an empirically-supported theory to define the cognitive abilities their instrument measures, (d) control for g, and (e) investigate whether the magnitude of gender differences varies developmentally if their sample represents a wide developmental span. The next section reviews the literature on gender differences in cognitive abilities and emphasizes the results from studies that meet these criteria.

1.2. Gender differences in cognitive abilities: an overview

Contemporary models of CHC theory propose the existence of at least seven broad cognitive abilities (Schneider & McGrew, 2012). The WPPSI-IV, the instrument used to investigate gender differences in this study, is designed to measure g and the following five broad cognitive abilities: comprehension-knowledge (Gc), visual processing (Gv), fluid reasoning (Gf), short-term memory (Gsm), and processing speed (Gs). For this reason, the current review of the gender differences literature is restricted to these five broad abilities and g, with special emphasis on young children.

1.2.1. Mean differences

Because of the power of general intelligence (g) to predict educational and occupational outcomes (Jensen, 1998), researchers have paid significant attention to gender differences in the mean of g. Studies that have investigated gender differences in g in children aged 5 to 17 using a latent variable approach generally support a null difference (Keith et al., 2011; Reynolds, Keith, Flanagan, & Alfonso, 2013) or an advantage for girls (Härnqvist, 1997; Reynolds et al., 2008; Rosén, 1995). Only a small number of studies offer information about gender differences in g in children younger than five, and these studies are limited in that they use an observed variable approach. Sellers, Burns, and Guyrke (2002) did not detect a gender difference in g in children aged 3 to 7 in the standardization sample of the WPPSI-R. In contrast, Burns and Reynolds (1988) discovered a gender difference in g favoring females aged 2 to 4 on the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983). In general, the research in children generally points to the absence of a gender difference in g or a female advantage.

At the level of the broad abilities, the mean gender difference that has received the most attention is the male advantage in Gv. Although a large volume of research supports a male advantage in Gv (Härnqvist, 1997; Keith et al., 2011; Reynolds et al., 2013; Reynolds et al., 2008; Rosén, 1995), the age at which this gender difference emerges is not evident, even when only considering studies using a latent variable approach. For example, one latent variable study suggests that the male advantage emerges at least by age 6 (Reynolds et al., 2008), whereas another latent variable study suggests that it does not emerge until age 18 (Keith et al., 2008). Studies using a latent variable approach to investigate gender difference in Gv have not included children younger than five. Other studies of young children using less robust methods have typically focused on observed scores of narrow Gv abilities. For this reason, more research that investigates

gender differences in Gv using a latent variable approach in children younger than five is needed to clarify when the male advantage emerges.

Another mean gender difference that has received significant attention is a proposed gender difference in Gc. Gc is the range and depth of knowledge that a person has acquired (Schneider & McGrew, 2012). Gc includes measures of language development, but the domain of Gc is broader because it can be assessed by tasks that require little to no expressive language (e.g. receptive vocabulary tests). The majority of studies using a latent variable approach indicate a male advantage in Gc in children aged 5 to 16 (Härnqvist, 1997; Keith et al., 2008; Reynolds et al., 2013; Reynolds et al., 2008; Rosén, 1995). Studies using a latent variable approach to investigate gender differences in Gc have not included children younger than five, and if the male advantage in Gc is a true difference, more research in young children is necessary to determine when it emerges.

A robust mean gender difference that has received less attention than the gender differences in g, Gv, or Gc is the female advantage in Gs. The female advantage in Gs emerges at least by age 5 (Camarata & Woodcock, 2006; Goldbeck, Daseking, Hellwig-Brida, Waldmann, & Petermann, 2010; Keith et al., 2008; Keith et al., 2011) and lasts across the lifespan (Härnqvist, 1997; Irwing, 2012). This finding emerges regardless of whether an observed or latent variable approach is used, although for an exception see Dolan et al. (2006). The magnitude of the standardized mean difference is typically at least 0.3.

Two cognitive abilities that generally do not demonstrate mean gender differences in children or adults are Gf and Gsm when studies use a latent variable approach (Dolan et al., 2006; Reynolds et al., 2013; Rosén, 1995). Keith et al. (2011) reported a difference in Gsm favoring girls aged 5 to 13 and a difference favoring boys aged 14 to 17. However, these findings are an exception, and studies that have investigated gender differences in Gf and Gsm in samples of children specifically 5 to 7 years of age have not found gender differences (Keith et al., 2008; Keith et al., 2011; Reynolds et al., 2008).

1.2.2. Variance differences

Researchers have historically paid more attention to gender differences in the mean of g and not the variance of g. However, males are overrepresented in populations that demonstrate very low cognitive ability (e.g. children with intellectual disabilities) and in populations that are assumed to demonstrate very high cognitive ability (e.g. Nobel Prize winners) (Dykiert, Gale, & Deary, 2009). For this reason, researchers have hypothesized that males demonstrate more variability in cognitive abilities than females. However, the support for the hypothesis of more variability in g in males is mixed. Studies that have used a latent variable methodology have generally failed to detect a gender difference in the variance of g (Härnqvist, 1997; Irwing, 2012; Keith et al., 2011; Reynolds et al., 2008; Rosén, 1995), whereas studies that have used an observed variable approach have found support for more male variability (Arden & Plomin, 2006; Calvin, Fernandes, Smith, Visscher, & Deary, 2010; Deary, Thorpe, Wilson, Starr, & Whalley, 2003; Dykiert et al., 2009; Strand, Deary, & Smith, 2006). Notably, the studies using an observed variable approach typically collect sample sizes even larger than the studies using a latent variable approach (n = 8700–320,000), so increased statistical power and a more adequate representation of the population may explain why observed variable studies have detected increased male variability in g.

The number of studies on gender differences in the variability of broad abilities is limited and their findings are inconsistent. Studies using a latent variable approach to answer this question have generally not detected gender differences in the variances of broad abilities (Keith et al., 2011), with the exception of Gsm. Reynolds et al. (2008) found more variability in Gsm in girls aged 12 to 14. However, they did not detect this gender difference in other children aged 6 to 12 and 14 to 18. Thus, additional research using a latent variable approach is needed to clarify discrepancies in the extant research regarding not only gender differences in the means of cognitive abilities but also in the variances of cognitive abilities.

1.3. The current study

The purpose of the present study is to investigate gender differences in the means and variances of g, Gc, Gv, Gf, Gsm, and Gs as measured by the WPPSI-IV in American children from 2 to 7 years of age. First, we used structural equation modeling to verify that the WPPSI-IV measures these abilities in the same way in males and females. Then gender differences in the means and variances these abilities were estimated. In addition, because the ages of 2 to 7 represent a wide developmental span, we examined whether the magnitude of gender differences varied across these ages.

2. Method

2.1. Measure

The Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition (WPPSI-IV) (Wechsler, 2012) is a norm-referenced test of cognitive abilities for children aged 2 years, 6 months to 7 years, 7 months (2:6–7:7). The WPPSI-IV is divided into two batteries of subtests. The battery of subtests for children from ages 2 years, 6 months to 3 years, 11 months (2:6–3:11) includes seven subtests that are designed to measure three CHC broad cognitive abilities (Gc, Gv, and Gsm) and g. The battery of subtests for children from ages 4 years, 0 months to 7 years, 7 months (4:0–7:7) includes 15 subtests that are designed to measure five CHC broad cognitive abilities (Gc, Gv, Gf, Gsm, and Gs) and g. Table 1 lists the names of the subtests, the broad ability that each measures, and their task demands. The internal consistency reliability coefficients for the Gc, Gv, Gf and Gsm subtests range from 0.82 to 0.95. The test–retest reliability coefficients for the Gs subtests range from 0.71 to 0.84.

The structure of both test batteries is based on CHC theory and is a second-order factor model. The subtests measure narrow cognitive abilities and are indicators of the broad abilities. The broad abilities are indicators of g. Based on confirmatory factor analyses by the test developers, for children aged 2:6 to 3:11, a second-order model with three first-order factors (Gc, Gv, and Gsm) and no residual covariances fits better than two alternative models. For children aged 4:0 to 7:7, a second-order model with five first-order factors (Gc, Gv, Gf, Gsm, and Gs) and no residual covariances fits better than five alternative models. These best fitting models were used in the study.

Table 1
WPPSI-IV subtest descriptions.

Subtest	Broad ability	Description
Information	Gc	Child answers questions that focus on general factual knowledge
Similarities	Gc	Child explains how two common objects or concepts are similar
Vocabulary	Gc	Child defines words
Comprehension	Gc	Child answers questions that focus on everyday problems and social situations
Receptive vocabulary	Gc	Child looks at pictures and selects the one the examiner names aloud
Picture naming	Gc	Child names objects
Block design	Gv	Child replicates a model or picture of designs using blocks within time limits
Object assembly	Gv	Child assembles puzzles of common objects within time limits
Matrix reasoning	Gf	Child views an incomplete grid and selects the missing portion from five options
Picture concepts	Gf	Child views an array of pictures and selects the ones that have similarities
Picture memory	Gsm	Child views one or more pictures for a limited time and then selects each one within a field of distracting pictures
Zoo Locations	Gsm	Child views one or more animal cards placed on zoo map for a limited time and then puts each card in the previously viewed location
Bug search	Gs	Child scans a group of bugs and marks the target bug within time limits
Cancellation	Gs	Child scans a random and nonrandom arrangement of pictures and marks target pictures within time limits
Animal coding	Gs	A key that pairs shapes and animals is presented. Child marks shapes that match presented animals within time limits.

2.2. Participants

The participants were the 1700 children in the normative sample of the WPPSI-IV. The normative sample is representative of the population of English-speaking children in the United States aged 2:6 to 7:7 based on age, gender, race/ethnicity, parent education level, and geographic region. The normative sample excluded children with language, visual, hearing, motor, medical, or psychological impairments that could depress test performance. Eleven hundred of the children were aged 2:6 to 3:11 and 600 of the children were aged 4:0 to 7:7. One-half of the children in each age group were male and one-half were female.

2.3. Analysis plan

The analyses were performed separately in the age 2:6–3:11 group and the age 4:0–7:7 group. The reason for this decision is that the factor model of the WPPSI-IV differs between these two age groups, and the procedures used in this study require imposing the same factor structure across participants. Two methods were used in this study: multiple group confirmatory factor analysis (MG-CFA) and multiple indicator-multiple cause (MIMIC) models. MG-CFA was used to investigate overall gender differences in the means and variances of g and the broad abilities in the two age groups. Because the magnitude of gender differences may vary by age within these groups, MIMIC models were used to investigate whether the magnitude of gender differences in the means of the cognitive abilities varied within each age group.

MG-CFA and MIMIC models are both methods for investigating the equivalence of the parameters of a factor model across multiple groups. In this study, the equivalence of the mean and variance parameters of the latent CHC broad abilities and g across females and males was of interest. MG-CFA and MIMIC models both can examine the equivalence of indicator intercepts and factor means of a CFA model across groups, but only MG-CFA can examine the equivalence of factor loadings, residual variances/covariances, and factor variances/covariances as well. One assumption of MIMIC models is that is the factor loadings, residual variances/covariances, and factor variances/covariances are the same across groups (Brown, 2006). The advantage of MG-CFA models is that they can test this assumption, while the advantage of MIMIC models is that testing the heterogeneity of factor means and indicator intercepts within a sample is simpler in the MIMIC approach and produces identical results (Reynolds et al., 2008).

A prerequisite to examining the equivalence of mean and variance parameters of factors in multiple groups is that the indicators measure the factors in the same way across groups. This condition is known as measurement invariance, and MG-CFA was used to test it. The prerequisite for comparing factor variances is that factor loadings are equivalent across groups (Brown, 2006). The prerequisite for comparing factor means is that factor loadings and indicator intercepts are the same across groups. In a second-order factor model, because the estimates of the first-order factors are theoretically error-free, the residual variances of the first-order factors are only composed of unique variance unexplained by the second-order factor (Chen, Sousa, & West, 2005). In a CHC model, the first-order factors represent the broad abilities. Therefore, the residual variances of the broad abilities represent a pure measure of their variance unexplained by g. Because the residual variances of the first-order factors were of substantive interest, the invariance of residual variances was also tested.

The equivalence of the parameters of a factor model is typically tested in a stepwise manner in which increasingly restrictive sets of equality constraints are placed on the parameters across groups. After each set of equality constraints is applied, if the fit of the model does not degrade significantly relative to a less restricted model, then the equality constraints are retained. Because the more restricted models are nested within the less restricted models, the fit of these models can be compared using a likelihood ratio test based on the chi-square (χ^2) difference between the more restricted model and the less restricted model.

Table 2 lists the steps used in this study for testing the invariance of the parameters of the factor models, which is based on the procedure used by Reynolds et al. (2008) to test the measurement invariance of a second-order factor model. Before estimating the factor models in males and females simultaneously, the models were estimated in males and females separately to verify that they fit the data of each

Table 2
Invariance testing procedure.

Models	Model specifications across sexes
0. Separate male and female models	Loading of most reliable subtest for each factor fixed to 1 + loading of Gf on g fixed to 1
1. Equal form	Same number of factors + same pattern of factor-indicator loadings + first-order factor intercepts and second-order factor mean fixed to zero for males and females
2. Equal subtest loadings on broad abilities	Model 1 + equal first-order factor loadings
3. Equal subtest intercepts	Model 2 + equal observed indicator intercepts + first-order factor intercepts freely estimated in females
4. Equal subtest residual variances	Model 3 + equal observed indicator residual variances
5. Equal broad ability loadings on g	Model 4 + equal second-order factor loadings
6. Equal broad ability residual variances and covariances	Model 5 + equal first-order factor residual variances + equal first-order factor covariances
7. Equal g variance	Model 6 + equal second-order factor variance
8. Equal broad ability means and g freely estimated	Model 7 + first-order factor intercepts fixed to zero for females + second-order mean freely estimated in females
9. Equal g mean	Model 8 + second-order mean fixed to 0 in females

group well. In the invariance testing process, the equal form model served as the baseline against which the more constrained models were compared.

Of note regarding the invariance testing procedure is that the equivalence of the broad ability means was tested by fixing the broad ability means to zero in males and females and freely estimating the mean of g (Table 2, Model 8). If the fit of the model degraded significantly, then this test indicated that a mean difference in g is not sufficient to account for mean differences across all subtests (Byrne & Stewart, 2006). The modification indices of the model were examined to determine which broad ability means needed to be freely estimated to significantly improve model fit. The broad ability means that had the largest modification indices were freed one at a time until the model fit comparably to the equal form model.

The absolute values of the factor means cannot be estimated in MG-CFA, but the differences between factor means can be estimated (Brown, 2006). The factor means are fixed to zero in one group, which is called the reference group. The factor means are freely estimated in the other group, and they represent the difference from the reference group's latent mean. Males were the reference group in this study. Consequently, positive differences indicated a female advantage and negative differences indicated a male advantage.

If the MG-CFA analyses indicated that a factor model met the assumptions for MIMIC analyses, MIMIC analyses were performed to determine whether the magnitude of gender differences varied by age. MIMIC models are referred to as CFA with covariates because the factors are regressed on observed variables, referred to as covariates. The covariates can be ordinal and represent group membership, or they can be continuous (e.g. age). A significant direct effect of a covariate on a factor or indicator signals that the factor or indicator mean differs depending on the value of the covariate. In this study, a covariate representing the interaction between gender and age covariates was of substantial interest. If the path between the gender–age interaction and a factor was significant, it signaled that the magnitude of the gender difference varied by age. To prevent multicollinearity between the covariates, the age covariate was centered.

The MIMIC analyses were performed in a stepwise manner. In the first model, gender, age, and the gender–age interaction covariates directly affected g. Non-significant paths from the covariates to g were deleted from the model. Then a direct effect of the gender, age, and gender–age interaction covariates on the broad abilities was specified. Because a model in which the three covariates directly affected the five broad abilities simultaneously would be underidentified, the effect of the three covariates on each broad ability was tested sequentially and non-significant effects were deleted.

Fit statistics used to evaluate the models were (a) χ^2, (b) Root Mean Square Error of Approximation (RMSEA), (c) Tucker Lewis Index (TLI), (d) Comparative Fit Index (CFI), (e) Standardized Root Mean Square Residual (SRMR), and (f) Akaike Information Criteria (AIC). Lower χ^2, RMSEA, SRMR, and AIC values and higher CFI and TLI values indicate better fitting models. RMSEA values below 0.06, CFI and TLI values from 0.95 to 1.0, and SRMR values less than 0.08 suggest good model fit (Hu & Bentler, 1999).

Normal theory maximum likelihood estimation was used to fit the models. One assumption of normal theory ML estimation is that the observed variables have a multivariate normal distribution. This assumption was examined by exploring whether each of the observed variables had a normal distribution. Kurtosis of the indicators ranged from 2.94 to 4.17 and skewness ranged from −0.36 to 0.33. Kurtosis and skewness values within these limits do not bias ML estimates of chi-square values (Curran, West, & Finch, 1996).

The input for the analyses was the raw data. Data were fully present for all observed variables. Test statistics were considered as significant if their probability was less than 5% ($p < .05$). MPlus version 7 (Muthén & Muthén, 2012) was used to estimate the MG-CFA and MIMIC models, and R version 3.0.2 (R Core Team, 2014) was used to extract the model fit indices and perform the chi-square difference testing.

3. Results

3.1. Ages 2:6–3:11

Table 3 presents descriptive statistics for the observed subtest and composite scores on the WPPSI-IV for children aged 2:6–3:11. The factor model of the WPPSI-IV proposed by the test developers for children aged 2:6–3:11 has three first-order factors and one second-order factor (see Fig. 1). The fit of this model was good in males and females based on the RMSEA, CFI, TLI, and SRMR fit statistics (see Table 4). The only fit index that indicated that this model fit poorly was χ^2 in males, which was significant. However, in large samples, χ^2 tends to be significant even when the differences between the observed and predicted covariances are slight (Brown, 2006). Because the other fit statistics indicated good fit of the factor model with three first-order factors and one second-order factor, this model was used in the invariance testing.

Table 3
WPPSI-IV index/subtest descriptive statistics ages 2:6–3:11.

Index/subtest	Female		Male			
	M	SD	M	SD	d	Variance ratio
Comprehension-knowledge index	102.64	14.68	97.27	14.86	0.37*	1.02
Information	10.43	2.94	9.37	3.07	0.36*	1.09
Receptive vocabulary	10.33	3.02	9.45	3.03	0.29*	1.01
Picture naming	10.44	2.87	9.61	3.02	0.29*	1.11
Visual processing index	100.84	15.26	97.39	13.97	0.23*	0.84
Block design	10.25	3.17	9.56	2.82	0.22*	0.8
Object assembly	9.94	3.1	9.44	3.03	0.16*	0.96
Short-term memory index	102.1	15.57	97.43	14.74	0.3*	0.9
Picture memory	10.33	3.14	9.51	3.18	0.26*	1.02
Zoo locations	10.3	3.05	9.55	2.96	0.25*	0.94
Full scale IQ	102.64	14.68	97.27	14.86	0.37*	1.02

Note. Variance ratios are calculated as male variance over female variance.
* $p < 0.05$, false discovery rate adjusted.

Table 5 provides the fit statistics for the models estimated in the invariance testing procedure for children aged 2:6–3.11. Equality constraints across gender on the first-order factor loadings, subtest intercepts, subtest residual variances, and second-order factor loadings did not significantly degrade model fit relative to the equal form model based on χ^2 difference testing (Models 2–5). Invariance of these parameters was required to meaningfully compare factor means and variances. Equality constraints on the broad ability unique variances and the variance of g did not significantly worsen model fit (Models 6–7), indicating that the variance of g and the broad abilities was not different for males and females. When the broad ability means were fixed to zero in males and females and g was freely estimated, model fit did not degrade significantly (Model 8). This result pointed to the absence of gender differences at the level of the broad abilities. However, constraining g to be equal across groups did worsen model fit significantly (Model 9), supporting the presence of a gender difference in g. The value of the standardized mean difference in g was 0.43, 95% CI [0.24, 0.61] and favored females. This difference translates to 6.41 units on an IQ scale (SD = 15).

Because the WPPSI-IV factor model for children aged 2:6–3:11 demonstrated measurement invariance across gender, a MIMIC model was implemented to investigate whether the magnitude of gender differences varied by age within this group. The effect of the gender–age interaction term on g or any of the broad abilities was not significant, indicating that the magnitude of gender differences on these abilities did not vary by age.

3.2. Ages 4:0–7:7

Table 6 presents descriptive statistics for the observed subtest and composite scores on the WPPSI-IV for children aged 4:0–7:7. The factor model of the WPPSI-IV proposed by the test developers for children aged 4:0–7:7 had five first-order factors and one second-order factor. The fit of this model was estimated for males and females separately before applying equality constraints to the models. The modification indices of this model for both genders suggested that estimating the covariance between the residual variances of the Gsm and Gs factors would considerably improve model fit. Specifying this covariance would be equivalent to specifying an intermediate factor between g and the broad abilities of Gsm and Gs. The presence of this intermediate factor is supported by factor analytic research on another test battery based on CHC theory (Taub & McGrew, 2013), and it is proposed to reflect information processing efficiency based on the cognitive performance model

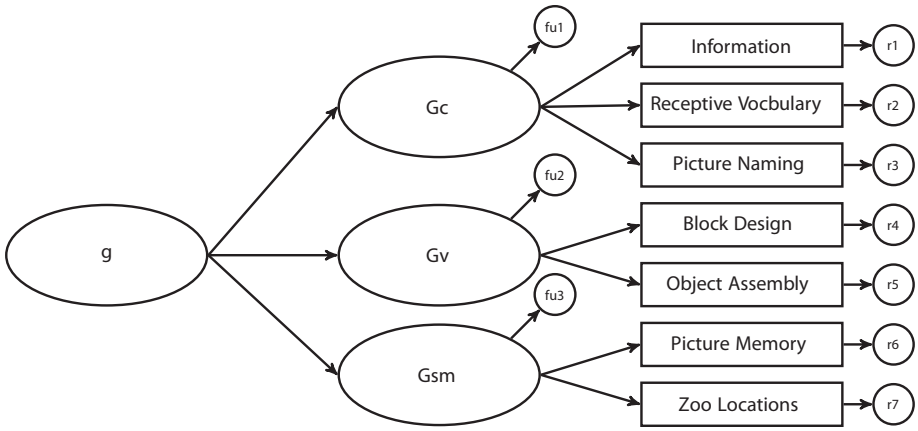

Fig. 1. Factor structure of the WPPSI-IV in children ages 2:6–3:11. This model was used for invariance testing.

Table 4
Fit statistics of baseline models in males and females separately.

Model	χ^2	df	p	RMSEA	CFI	TLI	SRMR
Females ages 4:0–7:7	7.59	11	0.75	0	1	1.01	0.02
Males ages 4:0–7:7	25.88	11	0.01	0.07	0.97	0.95	0.03
Females ages 2:6–3:11	146.03	84	<0.01	0.04	0.98	0.98	0.03
Males ages 2:6–3:11	211.01	84	<0.01	0.05	0.97	0.96	0.03

of abilities (Woodcock, 1993). Because a covariance between the Gsm and Gs factors is theoretically supported, it was estimated in subsequent male and female models. The fit of this updated model was good for males and females based on the CFI, TLI, and RMSEA fit statistics (see Table 4). As a result, this model was used in the invariance testing. Fig. 2 displays this model.

Table 7 provides the fit statistics for the models estimated in the invariance testing procedure for children aged 4:0–7.7. Equality constraints across gender on the first-order factor loadings, subtest intercepts, subtest residual variances, and second-order factor loadings did not significantly degrade model fit relative to the equal form model based on χ^2 difference testing (Models 2–5). Equality constraints on the broad ability unique variances and the variance of g did not significantly worsen model fit either, indicating that the variance of g and the broad abilities was not different for males and females (Models 6–7). When the broad ability means were fixed to zero in males and females and g was freely estimated, model fit degraded significantly (Model 8a). This result pointed to the presence of gender differences at the level of the broad abilities.

Based on the modification indices, freely estimating the mean of Gs was expected to produce the largest improvement in model fit. The decision to freely estimate the female mean of Gs is also well-supported by research indicating a gender difference in Gs. However, even when the means of g and Gs were freely estimated, the model continued to fit significantly worse than the equal form model (Model 8b). Another examination of the modification indices indicated that freely estimating the mean of Gv would produce the next largest expected improvement in model fit, another modification well supported by research. When the means of g, Gs, and Gv were freely estimated, the model fit was comparable to the fit of the equal form model (Model 8c). The value of the standardized mean difference on Gs was 0.21, 95% CI [0.09, 0.32] and favored females. This difference translates to 3.12 units on an IQ scale. The value of the standardized mean difference on Gv was −0.17, 95% CI [−0.28, −0.06] and favored males. This difference translates to 2.49 units on an IQ scale.

When the first-order intercepts of a second-order factor are not invariant, one could argue that the second-order factor mean cannot be compared meaningfully between groups because second-order mean differences may only reflect first-order intercept differences. In this case, the broad ability means are the first-order intercepts and were not invariant. Consequently, one may argue that the mean of g cannot be compared meaningfully across males and females. However, three of the five broad ability means (Gc, Gf, and Gsm) were equivalent and researchers have argued that factor means can be meaningfully compared when at least two of its indicators intercepts are invariant (Gregorich, 2006; Steenkamp & Baumgartner, 1998). This condition is called partial invariance (Byrne, Shavelson, & Muthén, 1989). For this reason, despite mean differences in the broad abilities, the significance of the mean difference in g was tested by constraining it to be equal in males and females. Constraining g to be equal across groups significantly degraded model fit (Model 9), pointing to the presence of a gender difference in g. The value of the standardized mean difference on g was 0.21, 95% CI [0.08, 0.34] and favored females. This difference translates to 3.15 units on an IQ scale.

Because the WPPSI-IV factor model for children aged 4:0–7:7 demonstrated measurement invariance across gender, a MIMIC model was implemented to investigate whether the magnitude of gender differences varied by age within this group. The effect of the gender–age interaction term on g or any of the broad abilities was not significant, indicating that the presence and magnitude of gender differences on these abilities did not vary by age.

4. Discussion

This study investigated gender differences in the means and variances of g and CHC broad abilities measured by the WPPSI-IV in children aged 2 to 7 years. From ages 2–3 years, the WPPSI-IV measures the CHC cognitive abilities of g, Gc, Gv, and Gsm. From ages 4–7 years, the WPPSI-IV measures g, Gc, Gv, Gf, Gsm, and Gs. Because the WPPSI-IV measures different

Table 5
Invariance testing model comparisons ages 2:6–3:11.

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	CFI	AIC
1. Equal form model	33.47	22				0.99	20,022.1
2. Equal subtest loadings on broad abilities	35.03	26	1.57	4	0.82	0.99	20,015.66
3. Equal subtest intercepts	36.85	30	3.38	8	0.91	1	20,009.48
4. Equal subtest residual variances	49.23	37	15.77	15	0.4	0.99	20,007.86
5. Equal broad ability loadings on g	52.82	39	19.35	17	0.31	0.99	20,007.45
6. Equal broad ability residual variances and covariances	55.12	42	21.65	20	0.36	0.99	20,003.75
7. Equal g variance	55.98	43	22.52	21	0.37	0.99	20,002.61
8. Equal broad ability means and g freely estimated	57.42	45	23.95	23	0.41	0.99	20,000.05
9. Equal g mean	78.12	46	44.65	24	0.01	0.98	20,018.75

Note: $\Delta\chi^2$ values are relative to the equal form model.

Table 6
WPPSI-IV index/subtest descriptive statistics ages 4:0–7:7.

Index/subtest	Female		Male			
	M	SD	M	SD	d	Variance ratio
Comprehension-knowledge index	101.01	14.53	99.03	15.41	0.14	1.12
Information	10.32	2.81	9.9	3.02	0.15*	1.15
Similarities	10.17	2.93	9.86	3.09	0.1	1.11
Vocabulary	10.15	2.92	9.8	3.15	0.12	1.16
Comprehension	10.33	2.92	9.6	3.12	0.25*	1.14
Receptive vocabulary	10.34	2.98	9.81	3.09	0.18*	1.08
Picture naming	10.21	3	9.89	3	0.11	1
Visual processing index	100.67	14.44	100.26	15.84	0.03	1.2
Block design	9.99	2.81	10.02	3.1	−0.01	1.22
Object assembly	10.15	3	9.97	3.1	0.06	1.06
Fluid reasoning index	100.88	14.6	99.04	15.44	0.13	1.12
Matrix reasoning	10.17	2.91	9.78	3.14	0.13	1.17
Picture concepts	10.14	2.99	9.9	3.1	0.08	1.08
Short-term memory index	101.37	14.58	98.87	14.98	0.17*	1.06
Picture memory	10.2	2.89	9.74	2.99	0.16*	1.08
Zoo locations	10.25	2.87	9.87	2.99	0.13	1.08
Processing speed index	102.29	15	97.68	14.77	0.31*	0.97
Bug search	10.27	3.03	9.65	3.02	0.21*	0.99
Cancellation	10.45	2.98	9.47	2.96	0.33*	0.98
Animal coding	10.38	3.03	9.56	2.98	0.27*	0.97
Full scale IQ	101.26	14.36	98.79	15.58	0.17*	1.18

Note. Variance ratios are calculated as male variance over female variance.
* $p < 0.05$, false discovery rate adjusted.

cognitive abilities in these two age groups, the analyses were performed separately for each group. MG-CFA was used to verify that the WPPSI-IV measured g and the broad abilities in the same way and estimate the magnitude of gender differences in the means and variances of the abilities. The MG-CFA analyses indicated that the WPPSI-IV measured the broad abilities and g in the same way across gender, which is a prerequisite to comparing their means and variances. The implication of this finding is that practitioners can make inferences about broad abilities based on WPPSI-IV subtest scores in the same way for males and females.

In the age 2–3 group, the mean of g was 0.43 standard deviations higher in females than in males. Gender differences in the means of Gc, Gv, and Gsm were not significant. In the age 4–7 group, g was 0.21 standard deviations higher in females, Gs was 0.21 standard deviations higher in females, and Gv was 0.17 standard deviations higher in males. Gender differences in the means of Gc, Gf, and Gsm were not significant in this age group. MIMIC models indicated that the magnitude of these gender differences did not vary within the two age groups. In both age groups, the variances of g or any of the broad abilities were not significantly different in males and females.

Gender differences in g have not been investigated using a latent variable approach in children younger than five, and the finding of a substantial female advantage in latent g in children as young as 2 years represents a novel contribution to the literature. Burns and Reynolds (1988) reported a standardized mean difference in g of 0.24 in children aged two to four using an observed variable approach. The magnitude of the standardized gender difference of g in the present study ($d = 0.43$) is larger, and the use of a latent variable approach may contribute to the more robust difference found in the present study.

The female advantage in g continued through ages 4–7 in this study, although the magnitude of this advantage decreased by half. Reynolds et al. (2008) reported a comparable standardized mean difference in latent g of 0.22 favoring females aged 6 to 8. However, two other latent variable studies have not reported gender differences in g in children aged 5 to 8 (Keith et al., 2008; Keith et al., 2011). The correlation between estimates of latent g from various cognitive ability tests based on CHC theory is very high (mean $r = .95$) (Floyd, Reynolds, Farmer, & Kranzler, 2013). Therefore, differences in test battery composition likely do not explain the discrepant results in these studies. The sample size for the age 4 to 7 group in the present study ($n = 1100$) is higher than sample size in other latent variable studies that have investigated a comparable age group ($n = 600$–800). Relatively higher power in the present study compared to past latent variable studies may contribute to the discrepant findings regarding gender differences in g.

One hypothesis for the female advantage in g is that in childhood, female brains mature earlier than do male brains. A longitudinal neuroimaging study of children has shown that total cerebral volume and total gray matter volumes peak one to two years earlier in girls than in boys (Lenroot et al., 2007), and gray matter volume and g are robustly correlated in pediatric populations (Taki et al., 2012; Wilke, Sohn, Byars, & Holland, 2003). However, studies directly linking gender dimorphism in cognitive abilities to gender dimorphism in brain structure and functioning in young children are lacking, and it is an area needing future attention.

The finding of a gender difference in Gs favoring females aged 4–7 further supports research indicating that this advantage emerges at least by preschool (Keith et al., 2008; Keith et al., 2011). Reviews of gender differences often neglect to discuss the female advantage in Gs (Hines, 2011; Miller & Halpern, 2014). However, given the consistency of this finding across the lifespan, its implications demand more attention. For example, the female advantage in Gs may contribute to the female advantage in general reading and writing skills that emerges as

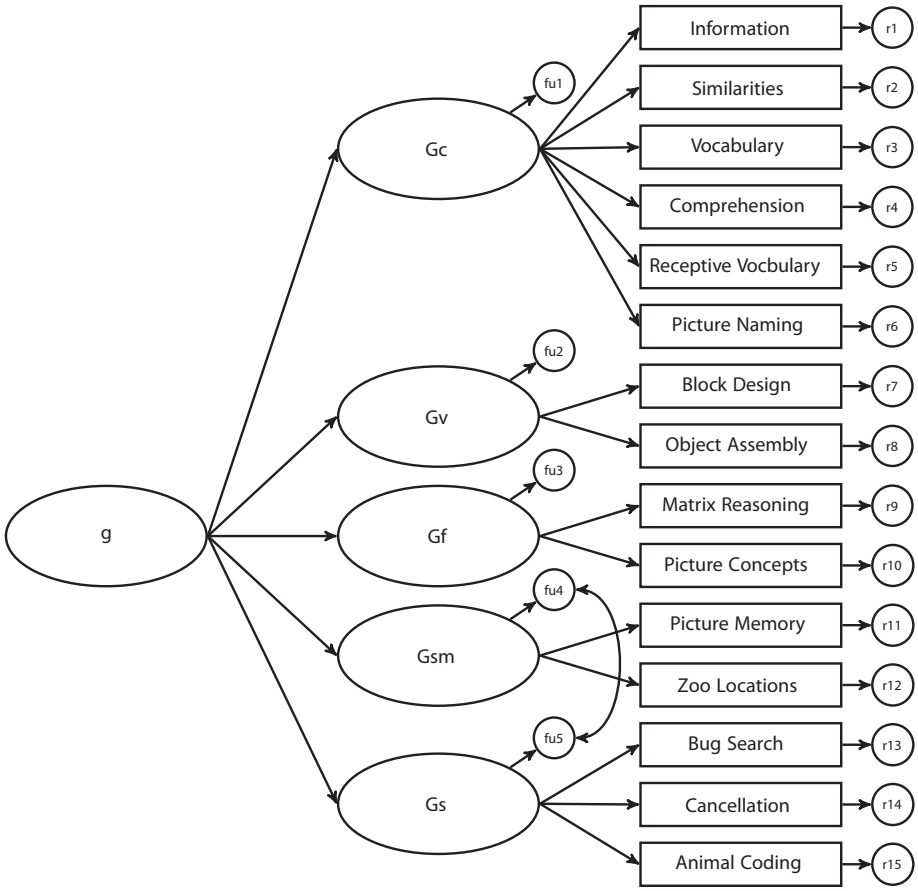

Fig. 2. Factor structure of the WPPSI-IV in children ages 4:0–7:7. This model was used for invariance testing.

early as kindergarten (Chatterji, 2006) and persists in adolescence (Stoet & Geary, 2013). Notably, girls demonstrate higher automaticity with basic reading and writing skills than boys (Camarata & Woodcock, 2006; Fearrington et al., 2014; Jewell & Malecki, 2005; Malecki & Jewell, 2003; Quinn & Wagner, 2013).

Gs has a direct and moderately sized effect beyond g on basic reading skills, basic writing skills, and written expression in children aged 6–13 (Floyd, McGrew, & Evans, 2008; McGrew & Wendling, 2010), supporting Gs as a candidate for explaining the gap between males and females in reading and writing. Theoretically, a relation between Gs and academic skills is expected because automaticity with basic cognitive skills allows students to allocate more cognitive resources to advanced aspects of academic tasks. Although gender differences in Gs likely do not fully explain gender differences in academic achievement, they may contribute to it. More research testing this hypothesis through a longitudinal design would advance the literature.

Boys aged 2–3 did not have an advantage in Gv but boys aged 4–7 did, which may indicate that the gender difference in

Table 7
Invariance testing model comparisons ages 4:0–7:7.

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	CFI	AIC
1. Equal form model	357.04	168				0.97	76,244.62
2. Equal subtest loadings on broad abilities	361.27	178	4.23	10	0.94	0.97	76,228.84
3. Equal subtest intercepts	382.97	188	25.93	20	0.17	0.97	76,230.54
4. Equal subtest residual variances	398.26	203	41.22	35	0.22	0.97	76,215.84
5. Equal broad ability loadings on g	401.7	207	44.66	39	0.25	0.97	76,211.27
6. Equal broad ability residual variances and covariances	413.04	213	56	45	0.13	0.97	76,210.62
7. Equal g variance	417.65	214	60.61	46	0.07	0.97	76,213.23
8a. Equal broad ability means and g freely estimated	444.63	218	87.59	50	<0.01	0.97	76,232.21
8b. g Gs mean differences freely estimated	427.73	217	70.69	49	0.02	0.97	76,217.3
8c. g Gs Gv mean differences freely estimated	418.92	216	61.88	48	0.09	0.97	76,210.49
9. Equal g mean	428.49	217	71.44	49	0.02	0.97	76,218.06

Note. $\Delta\chi^2$ values are relative to the equal form model.

Gv does not emerge until a minimum of age four. Levine, Huttenlocher, Taylor, and Langrock (1999) investigated gender differences on a measure of mental rotation, a Gv narrow ability, in children aged 4 to 7. Boys aged 4 years, 0 months to 4 years, 6 months did not have an advantage in Gv but boys age 4 years, 7 months and above did, supporting the theory that the Gv gender difference may not emerge until a minimum of age 4. However, some studies have detected a male advantage in mental rotation as early as infancy using a looking-time paradigm (Moore & Johnson, 2008; Quinn & Liben, 2008, 2014). In contrast, other studies have not found evidence for a gender difference in mental rotation in children aged 4 and older (Estes, 1998; Frick, Daum, Walser, & Mast, 2009; Frick, Hansen, & Newcombe, 2013; Platt & Cohen, 1981). These results do not support the theory that age 4 marks the emergence of the Gv gender difference. However, these studies used observed measures of a narrow Gv ability, which may explain the inconsistency in the literature. For this reason, more research modeling Gv as a latent variable is needed to determine the age at which the gender difference in Gv emerges.

The findings did not support a male advantage in Gc in either age group in the present study, which is inconsistent with multiple studies using a latent variable approach that have included children within the ages of 6 to 7 (Keith et al., 2008; Reynolds et al., 2013; Reynolds et al., 2008). However, a study by Keith et al. (2011) that used a latent variable approach did not detect a gender difference in Gc in children aged 5 to 8, a finding that is consistent with the current study. The Gc construct was represented by subtests that measured three different narrow Gc abilities (general verbal information, language development, and lexical knowledge) in each age group, and typically the composite of two narrow abilities is regarded as an adequate representation of a broad ability (Flanagan & McGrew, 1997). Therefore, construct underrepresentation is not a likely explanation for the failure to detect a gender difference in Gc. The lack a gender difference in Gc cannot be attributed to a difference at the narrow ability instead of the broad ability level either, as the Gc subtest intercepts were invariant across males and females.

Based on theories of cognitive development, a gender difference in Gc is not necessarily expected. One theory of the development of Gc is investment theory (Cattell, 1963). Cattell proposed that children acquire knowledge (Gc) by applying their fluid reasoning abilities to learning in different domains. For example, lexical knowledge develops by inferring the meaning of words in their context. Because research generally does not support a childhood gender difference in Gf, males and females have equivalent Gf to invest in acquiring knowledge and develop comparable Gc. In addition, Cattell (1963) theorized that educational experiences influence Gc more than other cognitive abilities, a claim supported empirically (Rindermann, Flores-Mendoza, & Mansur-Alves, 2010). Because the proportion of males and females enrolled in preprimary and primary school does not meaningfully differ in the United States (United Nations Children's Fund, 2014), the girls and boys in this sample would not be expected to differ in their educational attainment. For these reasons, the absence of a gender difference in Gc is consistent with theories of intellectual development.

The finding that children did not demonstrate a gender difference in Gf or Gsm is consistent with the majority of results from other latent variable studies that have controlled for g (Keith et al., 2008; Reynolds et al., 2013; Reynolds et al., 2008). Although all the broad abilities are strongly related to g, researchers have proposed especially close relationships between Gf and Gsm with g. Researchers have repeatedly discovered that g explains virtually 100% of the variance in Gf in higher-order models of cognitive test batteries (Keith, Fine, Taub, Reynolds, & Kranzler, 2006; Reynolds, Keith, Fine, Fisher, & Low, 2007), and others have reported extremely high correlations of working memory, a narrow Gsm ability, with g (Colom, Rebolloa, Palaciosa, Juan-Espinosa, & Kyllonen, 2004; Kyllonen & Christal, 1990). Studies using an observed variable approach have found gender differences in Gf and Gsm (Lynn & Irwing, 2004, 2008), but considering the especially close relationship between these abilities and g, the reason for these findings may be a failure to control for g. To explore the validity of this argument, MG-CFA models of the WPPSI-IV were constructed that did not specify a second-order g factor. Girls aged 2 to 3 demonstrated a significant advantage in Gsm. Girls aged 4 to 7 exhibited a significant advantage in Gf but not Gsm. These results partially support the argument that a failure to control for g may explain the findings in other studies of gender differences in Gf and Gsm.

4.1. Limitations

The evidence for gender differences in the variability of cognitive abilities is mixed, and the results of this study do not support their presence in children ages 2 to 7. However, the WPPSI-IV normative sample excluded children with severe language, visual, hearing, motor, medical, or psychological impairments, who are overrepresented at the low tail of the distribution of cognitive ability (American Psychiatric Association, 2013). In addition, one contributor to gender differences in the variability of cognitive abilities is that males are overrepresented at the low tail in the distribution of cognitive ability (Johnson, Carothers, & Deary, 2008). Therefore, the exclusion of children with these impairments could have restricted the variability of male cognitive ability in the normative sample, explaining the failure of this study to detect gender differences in the variability in cognitive abilities. Studies with larger sample sizes that report less restrictive exclusionary criteria have reported greater male variability in general intelligence (Calvin et al., 2010; Deary et al., 2003; Dykiert et al., 2009; Strand et al., 2006), perhaps as a result of higher power and/or more adequate representation of children with low ability.

Another limitation of this study is that a cross-sectional design was used to investigate developmental differences in gender differences. A stronger method of investigating this question would have been a longitudinal design to control for between child differences. A longitudinal design was not implemented because repeated measurements were not available for the children in the standardization sample of the WPPSI-IV.

CHC theory was the only theory used to operationalize the cognitive abilities measured by the WPPSI-IV, which reflects a limitation of this study because researchers have developed alternative taxonomies of cognitive abilities. For example, Johnson and Bouchard (2005) have proposed the verbal, perceptual, and image rotation (VPR) model of intelligence, and Das, Naglieri, and Kirby (1994) have proposed the planning, attention, simultaneous, and successive (PASS)

model of cognitive ability. Future studies should examine whether different models of cognitive abilities identify similar sex differences. Based on PASS theory, Bardos, Naglieri, and Prewett (1992) reported a sex difference favoring females in planning ability as measured by the Cognitive Assessment System (CAS). Planning ability as measured by the CAS corresponds with Gs based on factor analytic studies (Keith, Kranzler, & Flanagan, 2001), and a sex difference in Gs favoring females is well-established. Therefore, this finding shows how multiple theories can converge in identifying the same sex differences, lending further support for the existence of the differences.

Finally, the test development process could have had unknown effects on the results. The WPPSI-IV developers deleted items that were biased by gender, parental education, race, and/or ethnicity based on differential item functioning (DIF) (Wechsler, 2012). However, if a subtest constructed as unidimensional in reality measures a multidimensional construct, then deleting items based on DIF may mask true sex differences (Molenaar & Borsboom, 2013). For example, the unidimensional Information subtest of the WPPSI-IV is a general knowledge subtest. However, in adults, Lynn, Irwing, and Cammock (2002) discovered six dimensions underlying a general knowledge test. Notably, sex differences were not uniform across these dimensions. It is unknown but possible that general knowledge is truly a multidimensional construct in young children.

5. Conclusion

Despite these limitations, this study offers a novel contribution to the literature by studying gender differences in broad abilities at the latent variable level in children aged 2 to 3 years and expanding the emerging research in children aged 4 to 7 years. By at least age 2, a female advantage in g emerges, and by at least age 4, a female advantage in Gs and a male advantage in Gv emerge. Although more research needs to be conducted to confirm this study's findings, it indicates that certain gender differences may develop in early childhood and represent candidates for explaining differences in educational outcomes. Programs developed to narrow the gender gap in these abilities with the goal of reducing disparities in educational outcomes might more optimally be implemented in early childhood for maximum effect.

Acknowledgments

Standardization data from the Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (WPPSI-IV). Copyright © 2012 NCS Pearson, Inc. Used with permission. All rights reserved.

This research was supported by funding provided by the Michigan State University College of Education.

References

American Psychiatric Association (2013). *The diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, *41*(1), 39–48. http://dx.doi.org/10.1016/j.paid.2005.11.027.

Bardos, A.N., Naglieri, J.A., & Prewett, P.N. (1992). Gender differences on planning, attention, simultaneous, and successive cognitive processing tasks. *Journal of School Psychology*, *30*(3), 293–305. http://dx.doi.org/10.1016/0022-4405(92)90012-T.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Burns, C.W., & Reynolds, C.R. (1988). Patterns of sex differences in children's information processing with and without independence from g. *Journal of School Psychology*, *26*(3), 233–242. http://dx.doi.org/10.1016/0022-4405(88)90003-9.

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466. http://dx.doi.org/10.1037//0033-2909.105.3.456.

Byrne, B.M., & Stewart, S.M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 287–321. http://dx.doi.org/10.1207/s15328007sem1302_7.

Calvin, C.M., Fernandes, C., Smith, P., Visscher, P.M., & Deary, I.J. (2010). Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in England. *Intelligence*, *38*(4), 424–432. http://dx.doi.org/10.1016/j.intell.2010.04.005.

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, *34*(3), 231–252. http://dx.doi.org/10.1016/j.intell.2005.12.001.

Cattell, R.B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. http://dx.doi.org/10.1037/h0046743.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the early childhood longitudinal study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*(3), 489–507. http://dx.doi.org/10.1037/0022-0663.98.3.489.

Chen, F.F., Sousa, K.H., & West, S.G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 471–492. http://dx.doi.org/10.1207/s15328007sem1203_7.

Colom, R., Rebolloa, I., Palaciosa, A., Juan-Espinosa, M., & Kyllonen, P.C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, *32*(3), 277–296. http://dx.doi.org/10.1016/j.intell.2003.12.002.

Curran, P.J., West, S.G., & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. http://dx.doi.org/10.1037//1082-989X.1.1.16.

Das, J.P., Naglieri, J.A., & Kirby, J.R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence*. Boston, MA: Allyn & Bacon.

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*(6), 533–542. http://dx.doi.org/10.1016/s0160-2896(03)00053-9.

Dolan, C.V., Colom, R., Abad, F.J., Wicherts, J.M., Hessen, D.J., & van de Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence*, *34*(2), 193–210. http://dx.doi.org/10.1016/j.intell.2005.09.003.

Dykiert, D., Gale, C.R., & Deary, I.J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence*, *37*(1), 42–47. http://dx.doi.org/10.1016/j.intell.2008.06.002.

Estes, D. (1998). Young children's awareness of their mental activity: The case of mental rotation. *Child Development*, *69*(5), 1345–1360. http://dx.doi.org/10.2307/1132270.

Fearrington, J.Y., Parker, P.D., Kidder-Ashley, P., Gagnon, S.G., McCane-Bowling, S., & Sorrell, C.A. (2014). Gender differences in written expression curriculum-based measurement in third- through eighth-grade students. *Psychology in the Schools*, *51*(1), 85–96. http://dx.doi.org/10.1002/pits.21733.

Flanagan, D.P., & McGrew, K.S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories tests, and issues* (pp. 314–325) (1st ed.). New York, NY: Guilford Press.

Floyd, R.G., McGrew, K.S., & Evans, J.J. (2008). The relative contributions of the Cattell–Horn–Carroll cognitive abilities in explaining writing achievement during childhood and adolescence. *Psychology in the Schools*, *45*(2), 132–144. http://dx.doi.org/10.1002/pits.20284.

Floyd, R.G., Reynolds, M.R., Farmer, R.L., & Kranzler, J.H. (2013). Are the general factors from different child and adolescent intelligence tests the same? Results from a five-sample, six-test analysis. *School Psychology Review*, *42*(4), 383–401 (Retrieved from http://www.nasponline.org/publications/spr/abstract.aspx?ID=3839).

Frick, A., Daum, M.M., Walser, S., & Mast, F.W. (2009). Motor processes in children's mental rotation. *Journal of Cognition and Development*, *10*(1–2), 18–40. http://dx.doi.org/10.1080/15248370902966719.

Frick, A., Hansen, M.A., & Newcombe, N.S. (2013). Development of mental rotation in 3- to 5-year-old children. *Cognitive Development*, *28*(4), 386–399. http://dx.doi.org/10.1016/j.cogdev.2013.06.002.

Goldbeck, L., Daseking, M., Hellwig-Brida, S., Waldmann, H.C., & Petermann, F. (2010). Sex differences on the German Wechsler Intelligence Test for Children (WISC-IV). *Journal of Individual Differences*, *31*(1), 22–28. http://dx.doi.org/10.1027/1614-0001/a000003.

Gregorich, S.E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11), 78–94. http://dx.doi.org/10.1097/01.mlr.0000245454.12228.8f.

Härnqvist, K. (1997). Gender and grade differences in latent ability variables. *Scandinavian Journal of Psychology*, *38*(1), 55–62. http://dx.doi.org/10.1111/1467-9450.00009.

Hines, M. (2011). Gender development and the human brain. *Annual Review of Neuroscience*, *34*(1), 69–88. http://dx.doi.org/10.1146/annurev-neuro-061010-113654.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118.

Immekus, J.C., & Maller, S.J. (2010). Factor structure invariance of the Kaufman Adolescent and Adult Intelligence Test across male and female samples. *Educational and Psychological Measurement*, *70*(1), 91–104. http://dx.doi.org/10.1177/0013164409344491.

Irwing, P. (2012). Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences*, *53*(2), 126–131. http://dx.doi.org/10.1016/j.paid.2011.05.001.

Jensen, A.R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Jewell, J., & Malecki, C.K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, *34*(1), 27–44 (Retrieved from http://www.nasponline.org/publications/spr/abstract.aspx?ID=1778).

Johnson, W., & Bouchard, T.J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393–416. http://dx.doi.org/10.1016/j.intell.2004.12.002.

Johnson, W., Carothers, A., & Deary, I.J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, *3*(6), 518–531. http://dx.doi.org/10.1111/j.1745-6924.2008.00096.x.

Kaiser, S.M., & Reynolds, C.R. (1985). Sex differences on the Wechsler Preschool and Primary Scale of Intelligence. *Personality and Individual Differences*, *6*(3), 405–407. http://dx.doi.org/10.1016/0191-8869(85)90069-8.

Kaufman, A.S., & Kaufman, N.L. (1983). *Kaufman Assessment Battery for Children (K-ABC) administration and scoring manual*. Circle Pines, MN: American Guidance Service.

Keith, T.Z., Fine, J.G., Taub, G.E., Reynolds, M.R., & Kranzler, J.H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fourth Edition: What does it measure? *School Psychology Review*, *35*(1), 108–127 (Retrieved from http://www.nasponline.org/publications/spr/abstract.aspx?ID=1828).

Keith, T.Z., Kranzler, J.H., & Flanagan, D.P. (2001). What does the Cognitive Assessment System (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock–Johnson Tests of Cognitive Ability (3rd Edition). *School Psychology Review*, *30*(1), 89–118 (Retrieved from http://www.nasponline.org/publications/spr/abstract.aspx?ID=1598).

Keith, T.Z., & Reynolds, M.R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we have learned from 20 years of research. *Psychology in the Schools*, *7*, 635–650. http://dx.doi.org/10.1002/pits.20496.

Keith, T.Z., Reynolds, M.R., Patel, P.G., & Ridley, K.P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III Tests of Cognitive Abilities. *Intelligence*, *36*(6), 502–525. http://dx.doi.org/10.1016/j.intell.2007.11.001.

Keith, T.Z., Reynolds, M.R., Roberts, L.G., Winter, A.L., & Austin, C.A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales—Second Edition. *Intelligence*, *39*(5), 389–404. http://dx.doi.org/10.1016/j.intell.2011.06.008.

Kyllonen, P.C., & Christal, R.E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433. http://dx.doi.org/10.1016/S0160-2896(05)80012-1.

Lenroot, R.K., Gogtay, N., Greenstein, D.K., Wells, E.M., Wallace, G.L., Clasen, L.S., et al. (2007). Sexual dimorphism of brain developmental trajectories during childhood and adolescence. *NeuroImage*, *36*(4), 1065–1073. http://dx.doi.org/10.1016/j.neuroimage.2007.03.053.

Levine, S.C., Huttenlocher, J., Taylor, A., & Langrock, A. (1999). Early sex differences in spatial skill. *Developmental Psychology*, *35*(4), 940–949. http://dx.doi.org/10.1037//0012-1649.35.4.940.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*(5), 481–498. http://dx.doi.org/10.1016/j.intell.2004.06.008.

Lynn, R., & Irwing, P. (2008). Sex differences in mental arithmetic, digit span, and g defined as working memory capacity. *Intelligence*, *36*(3), 226–235. http://dx.doi.org/10.1016/j.intell.2007.06.002.

Lynn, R., Irwing, P., & Cammock, T. (2002). Sex differences in general knowledge. *Intelligence*, *30*(1), 27–39. http://dx.doi.org/10.1016/S0160-2896(01)00064-2.

Malecki, C.K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools*, *40*(4), 379–390. http://dx.doi.org/10.1002/pits.10096.

McGrew, K.S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. http://dx.doi.org/10.1016/j.intell.2008.08.004.

McGrew, K.S., & Wendling, B.J. (2010). Cattell–Horn–Carroll cognitive–achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools*, *47*(7), 651–675. http://dx.doi.org/10.1002/pits.20497.

Miller, D.I., & Halpern, D.F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences*, *18*(1), 37–45. http://dx.doi.org/10.1016/j.tics.2013.10.011.

Molenaar, D., & Borsboom, D. (2013). The formalization of fairness: Issues in testing for measurement invariance using subtest scores. *Educational Research and Evaluation*, *19*(2–3), 223–244. http://dx.doi.org/10.1080/13803611.2013.767628.

Moore, D.S., & Johnson, S.P. (2008). Mental rotation in human infants: A sex difference. *Psychological Science*, *19*(11), 1063–1066. http://dx.doi.org/10.1111/j.1467-9280.2008.02200.x.

Muthén, L.K., & Muthén, B.O. (2012). *MPlus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Platt, J.E., & Cohen, S. (1981). Mental rotation task performance as a function of age and training. *The Journal of Psychology*, *108*(2), 173–178. http://dx.doi.org/10.1080/00223980.1981.9915260.

Quinn, P.C., & Liben, L.S. (2008). A sex difference in mental rotation in young infants. *Psychological Science*, *19*(11), 1067–1070. http://dx.doi.org/10.1111/j.1467-9280.2008.02201.x.

Quinn, P.C., & Liben, L.S. (2014). A sex difference in mental rotation in infants: Convergent evidence. *Infancy*, *19*(1), 103–116. http://dx.doi.org/10.1111/infa.12033.

Quinn, J.M., & Wagner, R.K. (2013). Gender differences in reading impairment and in the identification of impaired readers: Results from a large-scale study of at-risk readers. *Journal of Learning Disabilities*. http://dx.doi.org/10.1177/0022219413508323 (Advance online publication).

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (Retrieved from http://www.R-project.org/).

Raiford, S.E., & Coalson, D.L. (2014). *Essentials of WPPSI-IV assessment*. Hoboken, NJ: John Wiley & Sons.

Reynolds, M.R., Keith, T.Z., Fine, J.G., Fisher, M.E., & Low, J.A. (2007). Confirmatory factor structure of the Kaufman Assessment Battery for Children—Second Edition: Consistency with Cattell–Horn–Carroll theory. *School Psychology Quarterly*, *22*(4), 511–539. http://dx.doi.org/10.1037/1045-3830.22.4.511.

Reynolds, M.R., Keith, T.Z., Flanagan, D.P., & Alfonso, V.C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology*, *51*(4), 535–555. http://dx.doi.org/10.1016/j.jsp.2013.02.003.

Reynolds, M.R., Keith, T.Z., Ridley, K.P., & Patel, P.G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, *36*(3), 236–260. http://dx.doi.org/10.1016/j.intell.2007.06.003.

Rindermann, H., Flores-Mendoza, C., & Mansur-Alves, M. (2010). Reciprocal effects between fluid and crystallized intelligence and their dependence on parents' socioeconomic status and education. *Learning and Individual Differences*, *20*(5), 544–548. http://dx.doi.org/10.1016/j.lindif.2010.07.002.

Rosén, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction*, *5*(1), 37–62. http://dx.doi.org/10.1016/0959-4752(95)00002-K.

Schneider, W.J., & McGrew, K.S. (2012). The Cattell–Horn–Carroll model of intelligence. In D.P. Flanagan, & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144) (3rd ed.). New York, NY: Guilford Press.

Sellers, A.H., Burns, W.J., & Guyrke, J. (2002). Differences in young children's IQs on the Wechsler Preschool and Primary Scale of Intelligence—Revised as a function of stratification variables. *Applied Neuropsychology*, *9*(2), 65–73. http://dx.doi.org/10.1207/S15324826AN0902_1.

Steenkamp, J. -B.E.M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. http://dx.doi.org/10.1086/209528.

Steinmayr, R., Beauducel, A., & Spinath, B. (2010). Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence*, *38*(1), 101–110. http://dx.doi.org/10.1016/j.intell.2009.08.001.

Stoet, G., & Geary, D.C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PLoS ONE*, *8*(3), e57988. http://dx.doi.org/10.1371/journal.pone.0057988.

Strand, S., Deary, I.J., & Smith, P. (2006). Sex differences in Cognitive Abilities Test scores: A UK national picture. *British Journal of Educational Psychology*, *76*(3), 463–480. http://dx.doi.org/10.1348/000709905x50906.

Taki, Y., Hashizume, H., Sassa, Y., Takeuchi, H., Asano, M., Asano, K., et al. (2012). Correlation among body height, intelligence, and brain gray matter volume in healthy children. *NeuroImage*, *59*(2), 1023–1027. http://dx.doi.org/10.1016/j.neuroimage.2011.08.092.

Taub, G.E., & McGrew, K.S. (2013). The Woodcock–Johnson Tests of Cognitive Abilities III's Cognitive Performance Model: Empirical support for intermediate factors within CHC theory. *Journal of Psychoeducational Assessment*. http://dx.doi.org/10.1177/0734282913504808 (Advance online publication).

United Nations Children's Fund (2014). *The state of the world's children 2014 in numbers: Every child counts — revealing disparities, advancing children's rights*. New York, NY: Author (Retrieved from http://www.unicef.org/publications/index_71829.html).

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition technical and interpretative manual*. San Antonio, TX: Psychological Corporation.

Wilke, M., Sohn, J. -H., Byars, A.W., & Holland, S.K. (2003). Bright spots: Correlations of gray matter volume with IQ in a normal pediatric population. *NeuroImage*, *20*(1), 202–215. http://dx.doi.org/10.1016/s1053-8119(03)00199-x.

Woodcock, R.W. (1993). An information processing view of Gf-Gc theory. *Journal of Psychoeducational Assessment monograph series: Woodcock–Johnson Psycho-Educational Assessment Battery—Revised* (pp. 80–102). Cordova, TN: Psychoeducational Corporation.