J. Intell. **2014**, *2*, 36-55; doi:10.3390/jintelligence2020036

OPEN ACCESS

Journal of
Intelligence
ISSN 2079-3200
www.mdpi.com/journal/jintelligence

Article

Sex Differences in Fluid Reasoning: Manifest and Latent Estimates from the Cognitive Abilities Test

Joni M. Lakin [1,*,†] and James L. Gambrell [2,†]

[1] Department of Educational Foundations, Leadership, and Technology, Auburn University, Auburn, AL 36849, USA

[2] ACT, Iowa City, IA 52243, USA; E-Mail: James.Gambrell@act.org

[†] These authors contributed equally to this work.

[*] Author to whom correspondence should be addressed; E-Mail: Joni.Lakin@auburn.edu; Tel.: +1-334-844-4930; Fax: +1-334-844-3072.

Received: 6 March 2014; in revised form: 19 May 2014 / Accepted: 3 June 2014 / Published: 12 June 2014

Abstract: The size and nature of sex differences in cognitive ability continues to be a source of controversy. Conflicting findings result from the selection of measures, samples, and methods used to estimate sex differences. Existing sex differences work on the Cognitive Abilities Test (CogAT) has analyzed manifest variables, leaving open questions about sex differences in latent narrow cognitive abilities and the underlying broad ability of fluid reasoning (*Gf*). This study attempted to address these questions. A confirmatory bifactor model was used to estimate *Gf* and three residual narrow ability factors (verbal, quantitative, and figural). We found that latent mean differences were larger than manifest estimates for all three narrow abilities. However, mean differences in *Gf* were trivial, consistent with previous research. In estimating group variances, the *Gf* factor showed substantially greater male variability (around 20% greater). The narrow abilities varied: verbal reasoning showed small variability differences while quantitative and figural showed substantial differences in variance (up to 60% greater). These results add precision and nuance to the study of the variability and masking hypothesis.

Keywords: sex differences; cognitive abilities; quantitative reasoning; STEM

1. Introduction

Sex differences in intelligence are a perennial interest in human abilities research. The rise of hierarchical theories of intelligence, most notably the Cattell-Horn-Carroll model [1]. have led to studies of sex differences at every level specificity. Sex differences research on general intelligence (*g*), "broad" abilities including fluid intelligence (*Gf*), as well as the more specific "narrow" abilities (including such things as verbal comprehension) have led to conflicting results in terms of the size and direction of differences [2–5]. The enduring controversy over sex differences has also led to a number of theories about why so many conflicting results are found.

Given that mean differences are for the most part trivial in size [4–6], especially for general and many broad abilities, the observation of disproportionately large numbers of males in the right and left tail of the ability distribution has led to the variability hypothesis. This hypothesis holds that males exhibit greater variation than females in many cognitive ability domains, which may explain their overrepresentation in the tails of ability distributions and creates the appearance of mean differences in incomplete or selected samples [7,8]. Some researchers have proposed explanatory theories for greater variability (such as a bimodal distribution with an excessively large left tail caused by higher rates of birth defects in males [8]), but descriptive analysis of variability differences is also critical when it comes to estimating the size of the effect to be explained [9]. Importantly, the mere existence of variability differences, regardless of their cause, could explain differential representation of males and females at the extremes of the distributions of many cognitive traits.

Another important hypothesis in explaining variations in the results of sex differences studies is the masking hypothesis which holds that the method of extracting ability estimates influences the magnitude of differences observed [10]. Specifically, the masking effect results from not partialling out the effects of general ability when estimating broad or specific ability factors. When the general (or higher-order) factor does not show sex differences, it washes out true differences in broad and specific abilities and underestimates those differences. Likewise, it can create differences in the higher-order factor that are really due to differences in the lower-order abilities that contribute to the estimation of the higher-order factor [2,4]. This hypothesis will be explored in more detail in a later section.

The focus on broad and, especially narrow abilities in addition to general intelligence is important because of the implications many feel that these abilities (and the sex differences they show) have for participation of men and women in various careers. In particular, there has been a strong interest in the specific/narrow abilities of quantitative reasoning, math and science aptitude, and mechanical reasoning because of their potential impact on highly valued science, technology, engineering, and mathematics (STEM) fields [11–13].

Of course, differences in general ability are also of interest for a variety of reasons. While researchers have reached a degree of consensus on sex differences in some broad and specific abilities (e.g., consistent advantages for females in processing speed [Gs] [14]; large differences favoring males in mechanical reasoning [15]), the magnitude of differences in general intelligence (*g*) still sparks significant debate and conflicting results. In estimating sex differences in *g*, the choice of tests in a battery, the age and selection of the sample, and the methods used to analyze the data all appear to impact results [3,16]. What researchers can agree on is that conflating *g* with broad abilities confuses the discussion of sex differences in both areas of research [2,3,10]. Therefore, choice of methodology to describe differences in *g* and broad abilities is important for the estimation of sex differences in both means and variances [3–5].

1.1. Competing Hypotheses and Impact of Methodology

In the effort to uncover the nature of sex differences, research has repeatedly shown that methodology matters. In their review of the literature and empirical findings, Steinmayr *et al.* [3] found that restricted sampling [16], selection of tests, and, in particular, the statistical methods used to analyze the constructs of interest can impact the research results.

From studies comparing various measurement models (manifest variables, latent bifactor models, latent hierarchical models, *etc.*), researchers have put forward the *masking hypothesis* [10]. The masking hypothesis concerns whether sex differences arise from broad abilities (used to estimate general ability) or from general ability itself. In some cases, analyses showed that broad ability differences were independent of *g*. In particular, Johnson and Bouchard [10] found that small or nonexistent differences in *g* washed out substantial sex differences in broad abilities when the two constructs were comingled. As a result, they found that many of the specific ability tests showed larger mean sex differences with *g* variance partialled out than in the manifest scores with *g* variance included. Their conclusion was that large mean differences in broad abilities were not related to differences in g and that, overall, there was a non-significant difference in *g*. Brunner, Krauss, and Kunter [17] argued for a similar approach to studying sex differences in mathematics achievement, where they found substantial sex differences in mathematical ability once the influence of a general factor was partialled out.

Although Johnson and Bouchard's and Brunner *et al.*'s findings are compelling, other research has not confirmed this finding. Specifically, Lemos *et al.* [15] found the opposite trend in their study, showing that the mean differences they detected in subtest scores on a reasoning battery were entirely explained by differences in *g* (differences around 2–4 points favoring males), with the exception of mechanical reasoning which showed large mean differences and numerical reasoning which showed small differences, both independent of differences in *g*. A key limitation of their study was that their battery consisted of only five subtests (compared to much larger and varied batteries in Johnson & Bouchard), and, furthermore, that one of these subtests was mechanical reasoning, one of the few reasoning domains to show very large male advantages, which may have skewed their general factor to favor males.

Importantly, although little research has addressed the masking hypothesis with respect to the variability hypothesis, Brunner *et al.* [9] showed in their study of achievement that masking *can* occur with variability differences and thus warrants study with ability test batteries. In their study, partialling out general achievement from specific mathematics and reading achievement showed that although general achievement and manifest mathematics achievement demonstrated substantial variability differences (Variance Ratio [VR] = 1.23 and 1.18, respectively, where a VR of 1.23 indicates that the males are 23% more variable than females), specific mathematics achievement did not show differences in variability once general achievement was partialled out (M-g). (Note that a variance ratio (calculated as the ratio of male variance to female variance) greater than 1.0 indicates that males were more variable than females. Feingold [18] suggested that a variance ratio of 1.10 or greater would be of practical importance on these types of tests.) This finding was replicated for the other achievement domains they observed (reading, science, and problem solving)—greater male variability was *only* observed for general achievement and not broad abilities when a nested latent model was used. In contrast, manifest variables for these other domains (with *g* and broad ability confounded) all showed substantially greater male variability.

1.2. Previous Research on Manifest and Latent Differences in Means and Variability

Given the volume of literature on the magnitude of sex differences in *g* and broad abilities [11,12], a general review is not given here. In this review, we focus on studies using large representative samples, broad assessment batteries, and preferably reporting manifest and latent estimates of sex differences in both means and variances in *g* and broad abilities.

A small number of studies have compared sex differences in latent and manifest models and found mixed results as to the impact of model selection (latent *vs.* manifest) on the size and nature of mean differences observed. Very few studies have compared the effects of latent *versus* manifest models on estimates of *variance* differences.

Irwing [19] studied an adult population (age 16–89) using the WAIS-III norm sample. Two latent models were applied and the results indicated that both the hierarchical and bifactor multi-group confirmatory factor analyses (MG-CFA) yielded comparable mean estimates. The manifest and latent estimates of mean differences in *g* were also similar ($d = 0.18$ and $d = 0.22$, respectively). See Table 1. Irwing only touched on the variability hypothesis in his discussion, but, in fact, his data shows interesting differences in the estimates of variance ratios from manifest to latent factors. The manifest variables show variability differences in *g* (VR = 0.86, surprisingly showing greater variability for *females*) while the latent *g* shows effectively no difference in variability (VR = 1.04). For the broad factors measured by WAIS-III, effects varied. Verbal Comprehension and Perceptual Organization were not much affected (VR around 1.0 for VC and 1.1 for PO), but Working Memory and Processing Speed demonstrated larger difference in the latent model than the manifest variables (VR 1.39 *vs.* 1.01, respectively, for WM; 0.65 *vs.* 0.88 for PS). In sum, latent models in some cases increased the size of VR estimates while other cases decreased the VR estimates. Irwing used his results to argue that the common observation of greater male variability is an artifact of manifest variables and that latent models will not show variability differences, but the surprising observation of greater *female* variance in his manifest variables calls into question the original data and whether there is something unusual about the battery of tests or the sample that reverses the typical observation of greater male variability.

Table 1. Compilation of Irwing's findings from WAIS-III.

Scale	**Cohen's *d* Effect Size**		**VR**	
	Manifest	**Latent (Bifactor)**	**Manifest**	**Latent (Bifactor)**
Full-Scale IQ (*g*)	0.18	0.22	86%	104%
Verbal Comprehension	0.23	NR	99%	103%
Perceptual Organization	0.22	NR	110%	114%
Working Memory	0.24	NR	101%	139%
Processing Speed	−0.31	−1.30	88%	65%

Note. NR = Not reported.

Despite Irwing's contention, variability differences have been found to persist in latent models. For example, Keith *et al.* [20] (2008) estimated the variance ratio for *g* to be 1.18 from the WJ-III norm sample (ages 2–90). In a separate study of the DAS norms sample (ages 2–17), Keith *et al.* [14] estimated the variance ratio for latent *g* to be 1.10, indicating that boys were 10% more variable than girls. See Table 2. Their estimate of the variance ratio for the latent *Gf* (1.55) was substantial. Both Keith *et al.* studies had large samples and lend strong evidence that variability differences are not solely an artifact of manifest variable models.

Table 2. Results from Keith *et al.* [20] (WJ-III, ages 6–59) and Keith *et al.* [14] (DAS (2nd ed.), ages 5–17).

	WJ-III				**DAS (2nd. ed.)**	
	d [a]		**VR** [b]		***d*** [a]	**VR** [c]
	Manifest	**Latent**	**Manifest**	**Latent**	**Latent**	**Latent**
g	NA	0.08	NA	1.18	0.03 [d]	1.10
Gf	−0.03	−0.35 **	1.20	NR	0.00 [d]	1.55
Gc	0.08	−0.14 **	1.11	NR	−0.12 [d]	1.05
RQ	0.14	−0.21 **	1.17	NR	--	--
Gv	0.00	−0.24	1.05	NR	−0.12	1.20
Gs	−0.29	0.40 **	1.03	NR	0.11*	1.20
Glr	0.08	NR	1.05	NR	0.15	1.10
Ga	0.03	−0.13	1.09	NR	--	--
Gsm	−0.06	−0.09	1.16	NR	0.04	0.85

Note. Average results across age groups. The estimates for ages 6–17 were similar to the overall results except Gf which showed mean differences of −0.25. [a] Positive differences favor girls; [b] Variance Ratios greater than 1.0 indicate greater variability for boys; [c] Estimated from graph in original article; [d] Only reported for age 5–8.

In their study of general achievement, Brunner *et al.* [9] also showed important (though mixed) differences between manifest achievement variables and those from a nested latent model. See Table 3. Specifically, mean differences in mathematics achievement grew when a latent model was applied (from *d* = .10 to .21), though reading achievement stayed about the same (*d* = -.36 to -.39). The general achievement factor showed near-zero mean differences, but substantially greater male variability (VR = 1.23), consistent with many studies of ability distributions. Brunner *et al.* did not report variance information for most of their latent variables, but their Mathematics Achievement factor did show diminished variance effect (VR = 1.19 in manifest and 1.02 in latent models).

Table 3. Brunner *et al.*'s PISA achievement results.

	d [a]		**VR** [b]	
	manifest	**Latent**	**manifest**	**latent**
General achievement	NA	0.01	NA	1.23
Mathematics achievement	0.10	0.21	1.19	1.02
Reading achievement	−0.36	−0.39	1.22	NR
Science achievement	0.05	NR	1.16	NR
Problem solving	−0.01	NR	1.18	NR

Note. [a] Positive differences favor girls; [b] VR > 1.0 indicates greater variability for boys.

Steinmayr *et al.* [3] analyzed a relatively small (N = 977) sample of students age 16–18 to compare the impact of model selection on sex differences. The assessment battery was the I-S-T 2000 R, which consists of nine reasoning tasks and a knowledge test. Because of their relatively small and nonrepresentative sample (coming from a university-track school), the exact estimates of differences themselves are not compelling. However, the differences between the manifest and latent estimates are of interest. Specifically, mean differences for the three broad abilities were smaller (reversing sign for verbal) in the latent model compared to the manifest estimates. See Table 4. The method of estimating the models did not meaningfully affect the VR estimates, which remained considerable (*i.e.*, greater than 1.10) for several of the factors.

Table 4. Compilation of Steinmayr *et al.*'s findings from I-S-T 2000 R.

	d		**VR**	
	Manifest	**Latent**	**Manifest**	**Latent**
V	−0.43	0.23	0.96	1.11
N	−0.81	−0.49	1.12	1.06
F	−0.50	−0.19	1.12	1.13
Gf	−0.62	−0.62	0.99	1.09
Gc	−0.78	−0.77	1.22	1.29

Note. Steinmayr *et al.*'s sample was quite small compared to other studies on the topic (female N = 551 and males N = 426). Positive differences in *d* favor males; VR > 1.0 indicates greater male variability.

These findings indicate that there is reason to believe that greater male variability can persist in latent models, and is not an artifact of manifest variables. Thus, the latent model evidence does not contradict the variability hypothesis for *g* and other broad abilities, despite Irwing's [19] contention. Clearly, however, the magnitude of those differences can vary, depending on the sample under study and the construct considered.

1.3. The Current Study

The Cognitive Abilities Test (CogAT) [21,22] is a battery of reasoning tasks measuring verbal, quantitative, and figural reasoning abilities for students in grades K-12 in the United States. Previous research on the CogAT [23,24] used observed (manifest) scores, which is appropriate when the goal is to inform practical uses of assessment results. In this study, our aim is to extend this analysis using latent models to probe sex differences at the construct level.

In this study, we considered *Gf*, three narrow abilities subsumed under *Gf*, and residual narrow abilities (with the general factor partialled out of the variance). Analyses were conducted on the norms samples from Form 6 and 7 of the CogAT. Previous work on the CogAT has not explored the general factor from the batteries or latent variables, but has addressed mean and variability differences in manifest variables representing the three reasoning batteries—Verbal, Quantitative, and Nonverbal (figural) Batteries—as measured in four cohorts between 1984 and 2011. Consistent with work on other batteries, the Nonverbal (figural) Battery showed negligible differences across test forms while the Quantitative Battery showed slight male advantages (0.05 to 0.15 across forms) and the Verbal Battery showed slight female advantages (−0.11 to −0.04). Variance ratios showed consistent advantages for males, with quantitative showing the largest differences (VR = 1.21 to 1.53). The greater differences in means and variance for quantitative reasoning is consistent with previous work [17,20], but differences for all three batteries were considerable (*i.e.*, greater than 1.10).

Data from the CogAT is relevant to this discussion because the test represents a balanced measure of fluid intelligence under Carroll's definition [25], which includes inductive, quantitative, and sequential reasoning components. The data is also informative because sampling is intentionally representative of the school-going population in the U.S., with large samples ranging across the 5–18 age group.

2. Experimental Section

This study relied on the national standardization data from the 2000 (Form 6) [22] and 2011 (Form 7) [21] editions of the CogAT. For simplicity, the forms are referred to as CogAT 6 and CogAT 7. For CogAT 6, the data for levels A–G of the test (administered to students in grades 3–11) were included. The primary battery (grades K-2) was excluded because it used a different set of tests that measure somewhat different abilities. Level H (grade 12) was also excluded because of the comparatively small sample size and use of college students to supplement the sample [26]. For CogAT 7, the naming convention and grade organization of test levels was altered and negated these issues (see Table 4 for grade-age-level correspondence [27]). However, for consistency with the other forms, the data for grades K-2 and 12 for CogAT 7 were omitted from this study.

The student samples used in the standardizations of CogAT 6 and 7 were drawn using a stratified random sample of public and private schools (including Catholic schools). The sampling units (school buildings) were sampled within strata defined by region of the country (four levels), school-district size (five levels), and school socioeconomic status (SES; five levels). Randomly selected schools within each stratification level were asked to participate. Around 400 schools were sampled for CogAT 6 and 250 were sampled for CogAT 7.

Within participating schools, all students in relevant grades were administered the test, with school administrators determining exclusion or accommodations for students with disabilities or limited English proficiency. Schools were asked to include all students who could meaningfully engage with the tasks [26,27]. English learners comprised 4.0% of the CogAT 6 sample and 2.8% of the CogAT 7 sample. Of the students classified as English learners, just a fraction (around 18% for CogAT 6 and 7) received accommodations. Students with learning disabilities (as defined by the school district) comprised 6.0% of the CogAT 6 sample and 7.0% of the CogAT 7 sample. In CogAT 6 and 7 data, 32% and 48% of these students, respectively, received at least one accommodation while taking the test.

For the analyses that follow, sample weights were used. These weights were based on the stratifying variables (region, size, and SES) to achieve a representative sample of U.S. schools. The students in the sample were found to be representative of that population according to federal data, though weights did not adjust for individual characteristics [26,27]. The sample sizes and ethnicity distributions at each test level in the two standardization samples are shown in Tables 5 and 6.

Table 5. Sample Sizes by Test Level for Cognitive Abilities Test (CogAT) Forms 6 and 7.

CogAT 6 (2000)		**CogAT 7 (2010)**	
Level (Grade)	**N**	**Level (Grade)**	**N**
A (3)	14,152	9 (3)	6141
B (4)	14,309	10 (4)	6120
C (5)	15,146	11 (5)	6555
D (6)	13,407	12 (6)	5601
E (7)	12,454	13/14 (7–8)	9669
F (8–9)	18,237	15/16 (9–10)	7912
G (10–11)	11,234	17/18 (11)	3295

Table 6. Percentages by Ethnicity for CogAT 6 and 7.

	Form (Year)	
Ethnicity	**CogAT 6 (2000)**	**CogAT 7 (2010)**
White (not Hispanic)	65.0	55.7
Black	16.3	13.3
Hispanic	11.5	17.6
Asian/Pacific Islander	3.6	5.3
Native American	2.5	4.8

2.1. Measures

CogAT was designed to measure the full range of reasoning abilities that define general fluid reasoning (*Gf*). Each form and level of the CogAT consists of a verbal, quantitative, and figural battery. The choice of batteries is supported by Carroll's [25] factor analytic work which showed that the *Gf* factor is defined by three reasoning abilities: (a) *sequential or deductive reasoning*—verbal, logical, or deductive reasoning; (b) *quantitative reasoning*—inductive or deductive reasoning with quantitative concepts; and (c) *inductive reasoning*—typically measured with figural tasks. These correspond roughly with the three CogAT batteries.

CogAT 6 and 7 used the same three subtests to measure verbal reasoning (Verbal Classification, Verbal Analogies, Sentence Completion) and the same three subtests to measure figural reasoning (Figure Classification, Figure Analogies, Figure Analysis [paper folding task]). These formats are classic psychometric formats. Slightly different collections of subtests were used for quantitative reasoning for the two forms. For CogAT 6, quantitative reasoning was measured with Quantitative Relations, Number Series, and Equation Building. Number Series is a classic format, with students identifying the next number in a patterned sequence (e.g., 2 4 6 …?). The other formats are not as common. Quantitative Relations requires students to identify which of two quantities or concepts (e.g., a quarter and a dollar) is greater. The final format, Equation Building, required students to combine a set of numbers and operations (e.g., 3 5 6 + *) to mathematically yield one of the answer choices.

For CogAT 7, Equation Building and Quantitative Relations were replaced because Quantitative Relations showed a degree of verbal and *Gc* loading and both formats were quite speeded [27]. These formats were replaced with Number Analogies (a format used by the British version of the CogAT (called the CAT) [28]) and a new format called Number Puzzles. The Number Analogies test applied the traditional analogy format to quantitative relationships (e.g., [2:4] [3:6] [4:?]). The Number Puzzles format required examinees to determine the numerical value(s) represented by one or more geometric shapes that will make one or more equations true (e.g., $\Delta + 3 = 10$, $\Delta = ?$). The new formats showed less verbal loading than the preceding formats [29] and were given in contexts that were less speeded than the formats they replaced. Within each battery, the number of items and their difficulty remained approximately the same across forms.

Items on each test form were developed through an extensive tryout process that included screening for difficulty, discrimination, and differential item functioning (DIF) using a Mantel-Haenszel procedure [26]. For CogAT 6, DIF analyses indicated that, across all levels, 3 verbal items favored males and 4 items favored females [27]; one quantitative item favored males (but only at one of four grade levels in which it appeared), none favored females; and no figural items showed DIF for males and females. Items showing moderate DIF were balanced to favor boys and girls. These items were never among the most challenging items for a given subtest and rarely among the easiest, making them unlikely to impact analyses of differences in population variance. For CogAT 7, only verbal items were found to show DIF with roughly equal numbers favoring males and females (8 and 7, respectively, across all levels; less than 1% of items).

Items within each battery were scaled to create a unidimensional, cross-grade scale for each battery independently. Both forms used a 1-PL IRT model (with fit statistics used in selecting items) to develop a unidimensional scale for each battery [26,27]. For the verbal, quantitative, and figural scores, K–R 20 reliabilities are typically around 0.95. Research has shown that scores on CogAT 6 correlate with IQ scores from individually administered ability tests about as well as the IQ scores from different individually administered tests correlate with each other [30,31].

2.2. Analysis

At each grade level, raw scores on all subtests were converted into normalized Z-scores. Previous studies on this topic have used age-based rather than grade-based divisions, so the sample was divided into 3 age groups: 8–10, 11–13, and 14–17. These age ranges roughly correspond to grades 3–5, 6–8, and 9–11 in the U.S. For each age group and battery, effect sizes (using Cohen's *d*) and variance ratios were calculated. Within each of these age groups, the correlation matrix for each gender on each form was extracted. These correlation matrices, along with means and variances for each gender, were then submitted to a multiple-group confirmatory factor analysis (CFA) using Mplus 6.1.

A bifactor CFA model with a single Gf factor and verbal, quantitative, and figural content factors (V, Q, and F respectively) was used as the basis for the latent variable analysis. See Figure 1. The only difference between the CogAT 6 and CogAT 7 models, aside from the two new quantitative subtests, was a small but significant cross-loading of Quantitative Relations on the Verbal Factor correlations were constrained to zero. Standard measurement invariance testing procedures were conducted on each group, but model fit was good and no unexpected problems were found. Therefore, reported results are based on a scalar invariance model. Subtest loadings, intercepts, and residuals were fixed across gender; and factor means and variances for females were set to zero and one. Factor means and variances for males were freely estimated. Preliminary results revealed that factor means were somewhat underidentified without additional constraints. Specifically, average mean gender differences across V, Q, and F would be pushed onto the Gf factor in some models but not others, for no obvious reason. To stabilize this problem, the sum of V, Q, and F factor means for males was constrained to zero. This forced net average differences to appear on the Gf factor, which is consistent with theoretical interpretations of Gf. No restrictions were placed on variances.

3. Results and Discussion

Complete CFA parameters and model fit are reported in Appendix A and B. Consistent with previous research on battery-level scores, we found small-to-negligible advantages to girls on the verbal subtests and to boys on the quantitative subtests. The figural subtests were the least consistent, with paper folding slightly favoring boys and figure classification favoring girls, particularly on CogAT 6. See Table 7. No trends across age groups were apparent.

Figure 1. Path Diagram of Bifactor Model for CogAT 7. Parameters freed in male model, parameters in female model fixed to M = 0, SD = 1.

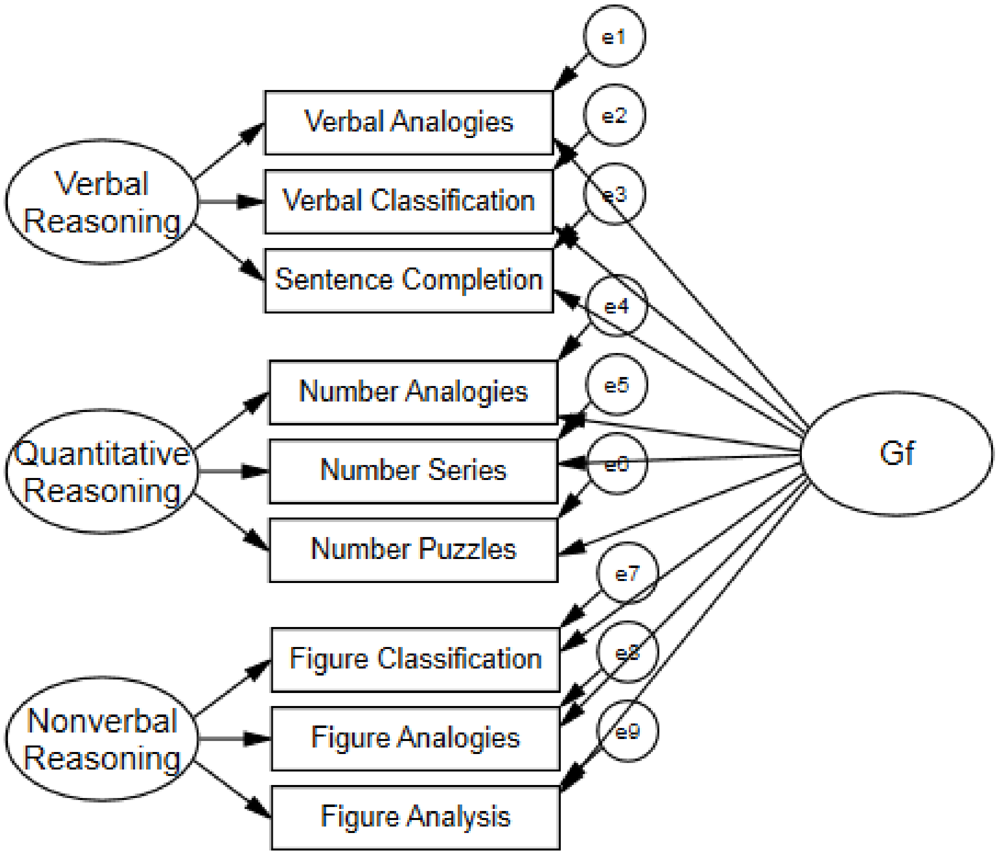

Table 7. Cohen's *d* effect size between male and female means by subtest, form, and age group.

	Age Group	VA	SC	VC	QR	NS	EB	FM	PF	FC
CogAT 6	8–10	−0.13	−0.13	−0.14	0.08	−0.01	−0.03	−0.06	0.09	−0.16
	11–13	−0.04	−0.09	−0.08	0.07	0.03	−0.07	−0.11	0.07	**−0.21**
	14–17	−0.08	−0.10	−0.09	0.08	0.06	−0.10	−0.13	0.09	**−0.20**
	Total	−0.10	−0.11	−0.08	0.08	0.02	−0.06	−0.09	0.08	**−0.19**
		VA	SC	VC	NA	NS	NP	FM	PF	FC
CogAT 7	8–10	−0.01	−0.12	−0.05	0.12	0.12	0.03	0.00	0.10	−0.08
	11–13	0.00	0.02	**−0.17**	0.14	**0.16**	0.06	0.03	**0.15**	**−0.15**
	14–17	0.01	0.05	−0.13	0.11	0.11	0.02	−0.10	0.09	−0.07
	Total	0.00	−0.03	−0.11	0.13	0.13	0.04	−0.02	0.12	−0.10
Cross-form	Total	−0.05	−0.07	−0.10	--[a]	0.08	--[a]	−0.05	0.10	−0.15

Note. *d* > 0 indicate higher male means. Effect sizes greater than 0.15 in bold. VA = Verbal Analogies, SC = Sentence Completion, VC = Verbal Classification, QR = Quantitative Relations, NS = Number Series, EB = Equation Building, FM = Figure Matrices, PF = Paper Folding (Figure Analysis), FC = Figure Classification, NA = Number Analogies, NP = Number Puzzles; [a] Subtests varied by form and cannot be averaged.

The variance ratios in Table 8 are also consistent with previous work. Almost every age group and subtest showed greater male variability (using a threshold of 1.1 as a meaningful level of difference [18]). A slight trend of increasing VRs with age on quantitative subtests may be present across both forms.

Table 8. Variance ratios (VRs) of males to females by subtest, form, and age group of CogAT.

	Age Group	VA	SC	VC	QR	NS	EB	FM	PF	FC
CogAT 6	8–10	**1.17**	**1.15**	**1.20**	**1.10**	**1.16**	**1.12**	**1.18**	1.08	**1.15**
	11–13	1.09	**1.23**	**1.15**	**1.16**	**1.19**	**1.19**	**1.27**	**1.11**	**1.21**
	14–17	1.02	**1.19**	**1.11**	**1.17**	**1.27**	**1.20**	**1.21**	**1.14**	**1.21**
	Total	**1.16**	**1.19**	**1.10**	**1.14**	**1.20**	**1.17**	**1.22**	**1.11**	**1.19**
		VA	SC	VC	NA	NS	NP	FM	PF	FC
CogAT 7	8–10	**1.16**	**1.12**	**1.10**	**1.14**	**1.27**	**1.13**	**1.19**	**1.15**	1.03
	11–13	1.03	1.04	**1.14**	**1.17**	**1.31**	1.08	**1.12**	**1.11**	**1.17**
	14–17	**1.12**	**1.21**	**1.25**	**1.26**	**1.40**	**1.12**	**1.19**	**1.20**	**1.27**
	Total	**1.10**	**1.11**	**1.15**	**1.18**	**1.31**	**1.11**	**1.17**	**1.15**	1.14
Cross-form	Total	**1.13**	**1.15**	**1.13**	--[a]	**1.25**	--[a]	**1.19**	**1.13**	**1.16**

Note. VR >1 indicate greater male variability. Values greater than 1.1 in bold. VA=Verbal Analogies, SC=Sentence Completion, VC=Verbal Classification, QR=Quantitative Relations, NS=Number Series, EB=Equation Building, FM=Figure Matrices, PF=Paper Folding, FC=Figure Classification, NA=Number Analogies, NP=Number Puzzles. [a] Subtests varied by form and cannot be averaged.

3.1. Latent Analyses—Mean Differences

Means and variances for latent factors are reported in Table 9. Model fit was excellent with CFIs above 0.988 and RMSEA confidence intervals below 0.05 for all age groups and forms. SRMR estimates averaged 0.023.

Latent models showed negligible mean differences in *Gf*. Because the domain factor means for males were constrained to sum to zero, any systematic difference across batteries is represented by a mean difference in *Gf*.

In general, the two forms showed consistent patterns of mean differences. The verbal factor favored girls somewhat, especially for younger groups (age 8–10 for CogAT 6, Ages 8–13 for CogAT 7). For the older age group, mean differences in verbal were negligible, though still favoring girls. On the quantitative factor, the two forms performed fairly consistently across age groups with moderate advantages favoring boys throughout. The figural factor again favored females, with larger differences at older groups that cancelled out the decrease in verbal means.

Table 9. Male Factor Means and Variances (Female values fixed to M = 0 SD = 1).

		Means				**Variances**			
Form	**Age Group**	**Gf**	**V**	**Q**	**F**	**Gf**	**V**	**Q**	**F**
CogAT 6	8–10	−0.06	**−0.22**	**0.27**	−0.04	**1.21**	**1.14**	0.95	**1.28**
	11–13	−0.05	−0.06	**0.25**	**−0.19**	**1.23**	1.06	**1.33**	**1.45**
	14–17	−0.06	−0.10	**0.29**	**−0.19**	**1.25**	0.99	1.06	**1.29**
	TOTAL	−0.06	−0.13	**0.27**	−0.14	**1.23**	1.06	**1.11**	**1.34**
CogAT 7	8–10	0.05	**−0.26**	**0.33**	−0.07	**1.23**	1.06	**1.55** [a]	**1.19**
	11–13	0.07	**−0.20**	**0.29**	−0.08	**1.17**	1.05	**1.60** [a]	1.07
	14–17	0.02	−0.07	**0.25**	**−0.18**	**1.37**	**1.11**	**1.26** [a]	**1.35**
	TOTAL	0.04	**−0.18**	**0.29**	−0.11	**1.25**	1.08	**1.47** [a]	**1.20**

Note. Positive means indicate higher factor scores for males. Values greater than 0.15 in bold. Variances > 1 indicate greater male variability. Values greater than 1.1 in bold. Standard errors for means are 0.024 or less. Male factor variances shown are comparable to Male/Female variance ratios. Standard errors for factor variances averaged 0.03, 0.04, 0.09, and 0.08 for Gf, V, Q, and F respectively. [a] Standard errors on CogAT 7 Quant were particularly large, averaging 0.15.

3.2. Differences in Variability

The *Gf* factor showed substantially greater male variability on both forms, with VRs ranging from 1.17 to 1.37. The median VR value of 1.2 implies a latent standard deviation difference of 10%, which has substantial implications for the tails of the ability distribution.

The verbal factor showed few meaningfully different variance ratios (*i.e.*, those over 1.1). These VRs are again quite similar to previous research and confirm that the variability hypothesis does not appear to apply to verbal reasoning, with or without partialling out the effects of *Gf*.

Results for the quantitative factor differed by form. CogAT 6 showed negligible variance differences with the exception of the 11–13 age group. Overall, variability for males appeared 11% greater on CogAT 6. In contrast, CogAT 7 showed the most substantial variability differences with males appearing on average 47% more variable (range 1.26–1.60). However, errors on the CogAT 7 estimates were quite large, indicating a weak quantitative domain factor. For example, the quantitative factor explained roughly 4.7% of the variance on the quantitative subtests in the 8–10 age group on CogAT 7. By contrast, variance explained by the domain factor was 16.4% on the verbal subtests and 12.6% on the figural at this age group. Variance explained by the general factor was 49.9%.

The figural factor also showed meaningful differences in variability, with CogAT 6 showing slightly greater disparities than CogAT 7. Compared to the quantitative factor, the CogAT 6 figural factor certainly showed greater variance differences than the quantitative factor, but the differences reversed for CogAT 7. Overall the variability differences ranged from 1.07 to 1.45 (median 1.28).

4. Conclusions

Mean sex differences in manifest scores from the subtests were consistent, on average, with previous work on manifest battery-level scores on the CogAT [23,24]. Interestingly, the latent mean differences were all larger (at least double, on average) than the manifest estimates. The biggest difference was for the quantitative factor, where manifest differences had a median of 0.06SD while the latent differences were around 0.28SD. Mean differences on *Gf*, which can only be estimated latently, were trivial, consistent with previous research showing no difference in *g* or *Gf* [5,14,32].

The results are also consistent with Johnson and Bouchard's [10] work, which showed that confounding broad and general ability variance leads to the underestimation of sex differences in broad and specific abilities (*i.e.*, the masking hypothesis). Importantly, our findings—that differences in reasoning reside in the more specific ability factors and not the general factor—contradict the results of Lemos *et al.* [15], who also studied a battery of reasoning abilities, and who found that differences in a general factor explained the differences in the specific test scores. The composition of their reasoning battery, particularly the inclusion of mechanical reasoning and spatial rotation, both of which strongly favor males, may explain the differences in our findings. Although the CogAT is not specifically designed to yield no mean differences between the sexes, the choice of reasoning tasks does omit any domains where sex differences are substantial.

The *Gf* factor showed substantially greater male variability on both forms, with VRs ranging from 1.17 to 1.37. The median value (1.23) was remarkably similar to the median effect size from CogAT 6 and 7 manifest scores in Lakin [24] which was 1.24 (across age groups and batteries). Though it was smaller than Keith *et al.*'s [14] estimate of a VR of 1.55 for *Gf*, our finding indicates that the variability hypothesis appears to apply equally to observed and latent score analyses, at least for *Gf*.

For the narrow abilities, partialling out *Gf* had different effects. For verbal, the small VRs in the manifest variables became mostly negligible in the latent analyses, indicating that verbal reasoning may be an exception to the frequent observation of greater male variability in quantitative and other domains (consistent with prior findings [18,33]). The figural factor was least affected, with similar (or slightly larger) VRs in the latent analyses.

Partialling out *Gf* impacted the quantitative battery the most, leaving a weakly measured residual specific ability with exaggerated differences in variability for CogAT 7. However, that latent residual factor still behaved consistently with prior research on the manifest Quantitative Battery scores, where VRs ranged from 1.21 to 1.53 across forms. VRs were substantially larger for CogAT 7, which may be a result of changes made to that battery compared to CogAT 6. These changes notably included replacing two speeded tasks (one of which was verbally loaded) with less speeded and more purely quantitative tasks. Quantitative reasoning may be the most sensitive to greater male variability, consistent with the conclusions of Mackintosh [4]. Interestingly, our findings contradicted Brunner *et al.* [17], who found that variability differences in their mathematics achievement factor disappeared in the latent analysis. Given the substantial differences in the assessment batteries across the two studies, it is unclear which results are more generalizable.

Differences in variance, even in the absence of mean differences, have important implications in practice. The *Gf* factor showed substantially greater male variability on both forms, with a median VR of 1.23. Importantly, a VR of 1.20 implies a latent standard deviation difference of 10%, which has substantial implications for the tails of the ability distribution. Using a normal distribution, such a difference in standard deviation would yield a male-female ratio of around 3:2 in the top 2% and a ratio of 5:2 in the top 0.2% (using cutscores based on female SDs). As previous work has shown [24,34], such ratios have been observed in studies of the extreme right tails of cognitive ability distributions and may have implications for why we observe relatively few women participating in elite levels of many math, science and engineering fields.

4.1. Limitations

One serious concern in interpreting the results for the three narrow abilities in this study is that once *Gf* is partialled out, we cannot be certain whether the factor variance that remains should be interpreted as verbal, quantitative, and figural reasoning; if it reflects a more specific trait; or is simply a method factor. The quantitative factor, in particular, seems to be fairly unreliable, with the largest standard errors, and thus should be interpreted with the most caution.

Another limitation of this study, and certainly all studies of this topic, is the choice of measures of the intended constructs, which can create or eliminate sex differences depending on the choice of tasks [2]. For example, although, in general, researchers agree that females show some advantage in verbal domains, other research has found that males are advantaged on some specific formats, such as verbal analogies [35,36]. The inclusion of verbal analogies in the CogAT may therefore diminish observed differences in verbal reasoning. Likewise, the omission of reasoning domains that strongly favor males (mechanical reasoning, spatial rotation [37]) may explain differences in our findings from Lemos *et al.* [15]. However, the selection of tasks on CogAT is consistent with the Cattell-Horn-Carroll theory of intelligence and definition of the *Gf* factor and is therefore defensible if not definitive. Future research might explore the use of indicators of the *Gf*-related narrow abilities (inductive, deductive, and quantitative reasoning) that do not confound content with task. The alignment of CogAT scales with the Berlin Model of Intelligence Structure [38] (which explicitly models verbal, quantitative, and figural facets) should also be explored.

Another limitation is the assumption of normality, which is inherent in latent analyses. This assumption makes detailed analysis of the tails of the distribution impossible in latent distributions. Previous work has shown substantial and important differences in the ratio of males to females at the extremes of the ability distribution on the CogAT batteries. The differences in variability observed in this study likely indicate similar differences in ratios in the latent variables, but cannot be directly tested. However, it is also quite possible that the true distribution of latent ability is non-normal in such a way as to create different pattern of ratios at the tails of the distributions [8].

Finally, the age range of our study did not permit us to directly test Lynn's hypothesis [39] that sex differences do not fully manifest themselves until early adulthood. However, there is no indication in our data of a trend of increasing differences, even at the oldest group which is close to the age at which

Lynn contends differences will manifest. A competing explanation that deserves future attention is the nature of sample recruitment. Dykiert *et al.* [16]. demonstrated that volunteer effects in norming samples for intelligence tests can be problematic for estimating means and variability. They show evidence that women are more likely to volunteer than men for testing as adults and that more intelligent individuals are more likely to volunteer. In their study, the combination of these two effects resulted in a greater range of women volunteering and, as a result, men showing proportionately larger advantages in IQ and somewhat greater variability as the volunteer effect was introduced. Dykiert *et al.* argue that these volunteer effects are much smaller for children, because participation in testing programs usually occurs in schools where there is little opportunity for students to opt out of testing. Therefore, one explanation for the lack of sex differences in *Gf* in our study is that volunteer effects do not impact our data as it does studies of adult samples.

4.2. Final Comments

This study weighs in on a number of hypotheses related to the nature of sex differences in broad and narrow/specific cognitive abilities. First, Lemos *et al.* [15] argue that differences in mechanical reasoning and similar domains are the key to differences in STEM engagement. While this may be true in general, this study shows that we cannot rule out differences in the variability of *Gf* as an additional explanatory factor for why males and females differ in their engagement in elite levels of STEM fields. Further, it suggests that variability differences could be explored in studies of sex differences in elite performance in other domains.

This research also shows, quite compellingly, that the *variability hypothesis* [8,18] is plausible and impacts both manifest and latent analyses of general ability. The *masking hypothesis* [10] was also supported for factor means. All three batteries showed greater mean difference in latent *vs.* manifest estimates while there were no substantial differences in the general ability factor. Further research is needed to explore whether the masking hypothesis may also apply to variability differences. Both the quantitative and figural factors showed some evidence of larger variability differences with *Gf* partialled out, while the verbal factor showed overall less variability with *Gf* partialled out. Unlike mean differences, there are clearly variability differences in *Gf* between males and females. When manifest variables are studied, the greater male variability in *Gf* is bound to inflate the variability in narrow abilities. This should be taken into consideration when selecting statistical models in future studies of sex differences in means and variability.

Acknowledgments

The authors thank David Lohman for providing the data for this manuscript and Joann Moore and Paul Westrick for their comments on the manuscript.

Author Contributions

Joni M. Lakin contributed to the study conception, analysis and interpretation of data, and was the primary author of the manuscript. James L. Gambrell contributed to the study conception, took the lead in analysis and interpretation of data, and contributed to drafting the manuscript.

Appendix

Appendix A. Model parameters and fit for CogAT 6.

		8–10		**11–13**		**14–18**	
	Parameter	**Est.** [a]	**S.E.**	**Est.** [a]	**S.E.**	**Est.** [a]	**S.E.**
Gf factor loadings	VA	0.74	0.01	0.72	0.01	0.71	0.01
	VC	0.57	0.01	0.62	0.01	0.63	0.01
	SC	0.68	0.01	0.65	0.01	0.64	0.01
	QR	0.73	0.01	0.73	0.01	0.74	0.01
	NS	0.77	0.01	0.77	0.01	0.77	0.01
	EB	0.74	0.01	0.71	0.01	0.71	0.01
	FC	0.73	0.01	0.72	0.01	0.74	0.01
	FM	0.78	0.01	0.79	0.01	0.77	0.01
	PF	0.66	0.01	0.68	0.01	0.68	0.01
V factor loadings	VA	0.38	0.01	0.47	0.01	0.49	0.01
	VC	0.29	0.01	0.42	0.01	0.48	0.01
	SC	0.50	0.01	0.52	0.01	0.53	0.01
	QR	0.15	0.01	0.16	0.01	0.15	0.01
Q factor loadings	QR	0.57	0.03	0.50	0.02	0.44	0.02
	NS	0.08	0.01	0.12	0.01	0.17	0.01
	EB	0.11	0.01	0.16	0.01	0.14	0.01
N factor loadings	FC	0.21	0.01	0.23	0.01	0.24	0.01
	FM	0.42	0.01	0.36	0.01	0.37	0.01
	PF	0.19	0.01	0.19	0.01	0.25	0.01
Intercepts	VA	0.06	0.01	0.03	0.01	0.05	0.01
	VC	0.05	0.01	0.03	0.01	0.04	0.01
	SC	0.08	0.01	0.03	0.01	0.05	0.01
	QR	−0.04	0.01	−0.04	0.01	−0.03	0.01
	NS	0.01 [NS]	0.01	0.01 [NS]	0.01	0.00	0.01
	EB	0.01 [NS]	0.01	0.00 [NS]	0.01	0.00	0.01
	FC	0.03	0.01	0.04	0.01	0.05	0.01
	FM	0.03	0.01	0.05	0.01	0.06	0.01
	PF	0.02	0.01	0.04	0.01	0.05	0.01
Residual variances	VA	0.24	0.00	0.19	0.00	0.18	0.00
	VC	0.55	0.01	0.39	0.00	0.32	0.00
	SC	0.22	0.01	0.24	0.00	0.24	0.00
	QR	0.07 [NS]	0.04	0.12	0.02	0.16	0.02
	NS	0.34	0.00	0.33	0.00	0.31	0.00
	EB	0.39	0.00	0.41	0.00	0.41	0.01
	FC	0.37	0.00	0.36	0.00	0.32	0.00
	FM	0.13	0.01	0.14	0.01	0.16	0.01
	PF	0.48	0.00	0.44	0.00	0.40	0.01
Model Fit [b]	ChiSq (df)	1580 (64)		2273 (64)		1599 (64)	
	sig.	<0.001		<0.001		<0.001	
	CFI	0.993		0.991		0.989	
	TLI	0.992		0.99		0.988	
	AIC	687651		734982		413451	
	RMSEA	0.037		0.043		0.047	
	(CI)	(0.035–0.038)		(0.041–0.044)		(0.045–0.049)	
	SRMR	0.017		0.021		0.026	

Note. [a] if not noted, $p < 0.001$. NS = non-significant. [b] model fit for male/female multiple-group model.

Appendix B. Model parameters and fit for CogAT 7.

		8–10		**11–13**		**14–18**	
	Parameter	**Est.** [a]	**S.E.**	**Est.** [a]	**S.E.**	**Est.** [a]	**S.E.**
Gf factor loadings	VA	0.71	0.01	0.70	0.01	0.65	0.01
	VC	0.67	0.01	0.61	0.01	0.61	0.01
	SC	0.63	0.01	0.63	0.01	0.61	0.01
	NA	0.68	0.01	0.72	0.01	0.69	0.01
	NS	0.70	0.01	0.69	0.01	0.68	0.01
	NP	0.69	0.01	0.73	0.01	0.69	0.01
	NS	0.70	0.01	0.69	0.01	0.68	0.01
	FC	0.66	0.01	0.61	0.01	0.61	0.01
	FM	0.69	0.01	0.63	0.01	0.62	0.01
	PF	0.60	0.01	0.63	0.01	0.61	0.01
V factor loadings	VA	0.27	0.01	0.38	0.01	0.43	0.01
	VC	0.40	0.01	0.48	0.01	0.48	0.01
	SC	0.48	0.01	0.51	0.01	0.55	0.01
Q factor loadings	NA	0.25	0.02	0.31	0.02	0.32	0.01
	NS	0.20	0.02	0.23	0.01	0.30	0.01
	NP	0.08	0.01	0.17	0.01	0.20	0.01
N factor loadings	FC	0.12	0.01	0.20	0.01	0.25	0.01
	FM	0.55	0.06	0.36	0.02	0.24	0.01
	PF	0.15	0.02	0.24	0.02	0.26	0.02
Intercepts	VA	0.02$^{<0.05}$	0.01	0.02 $^{\text{NS}}$	0.01	0.01 $^{\text{NS}}$	0.01
	VC	0.04	0.01	0.03$^{<0.01}$	0.01	0.01 $^{\text{NS}}$	0.01
	SC	0.05	0.01	0.03$^{<0.01}$	0.01	0.01 $^{\text{NS}}$	0.01
	NA	−0.06	0.01	−0.07	0.01	−0.05	0.01
	NS	−0.05	0.01	−0.06	0.01	−0.04	0.01
	NP	−0.03	0.01	−0.05	0.01	−0.03	0.01
	FC	−0.01 $^{\text{NS}}$	0.01	−0.01 $^{\text{NS}}$	0.01	0.02 $^{\text{NS}}$	0.01
	FM	0.00 $^{\text{NS}}$	0.01	−0.01 $^{\text{NS}}$	0.01	0.02 $^{\text{NS}}$	0.01
	PF	−0.01 $^{\text{NS}}$	0.01	−0.01 $^{\text{NS}}$	0.01	0.02$^{<0.05}$	0.01
Residual variances	VA	0.37	0.01	0.32	0.01	0.31	0.01
	VC	0.34	0.01	0.35	0.01	0.32	0.01
	SC	0.31	0.01	0.30	0.01	0.24	0.01
	NA	0.40	0.01	0.29	0.01	0.31	0.01
	NS	0.40	0.01	0.41	0.01	0.35	0.01
	NP	0.45	0.01	0.39	0.01	0.39	0.01
	FC	0.51	0.01	0.55	0.01	0.49	0.01
	FM	0.13$^{<0.05}$	0.06	0.42	0.02	0.47	0.01
	PF	0.57	0.01	0.51	0.01	0.48	0.01
Model Fit [b]	ChiSq (df)	685 (65)		1352 (65)		1088 (65)	
	sig.	<0.001		<0.001		<0.001	
	CFI	0.992		0.984		0.988	
	TLI	0.991		0.982		0.987	
	AIC	366028		366376		361033	
	RMSEA	0.033		0.048		0.042	
	(CI)	(0.031–0.035)		(0.045–0.050)		(0.040–0.045)	
	SRMR	0.019		0.029		0.025	

Note. [a] if not noted, $p < 0.001$. NS = non-significant. [b] model fit for male/female multiple-group model.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Schneider, W.J.; McGrew, K. *Contemporary Intellectual Assessment: Theories, Tests, and Issues*; Flanagan, D., Harrison, P., Eds.; Guilford: New York, NY, USA, pp. 99–144.
2. Mackintosh, N.J. Reply to Lynn. *J. Biosoc. Sci.* **1998**, *30*, 533–539.
3. Steinmayr, R.; Beauducel, A.; Spinath, B. Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence* **2010**, *38*, 101–110.
4. Mackintosh, N.J. *IQ and Human Intelligence*; Oxford University Press: Oxford, UK, 2011.
5. Jensen, A.R. *The g Factor: The Science of Mental Ability*; Praeger: Westport, CT, USA, 1998.
6. Hyde, J.S. The gender similarities hypothesis. *Am. Psychol.* **2005**, *60*, 581.
7. Feingold, A. Cognitive gender differences are disappearing. *Am. Psychol.* **1988**, *43*, 95–103.
8. Johnson, W.; Carothers, A.; Deary, I.J. Sex differences in variability in general intelligence: A new look at the old question. *Perspect. Psychol. Sci.* **2008**, *3*, 518–531.
9. Brunner, M.; Gogol, K.M.; Sonnleitner, P.; Keller, U.; Krauss, S.; Preckel, F. Gender differences in the mean level, variability, and profile shape of student achievement: Results from 41 countries. *Intelligence* **2013**, *41*, 378–395.
10. Johnson, W.; Bouchard, T.J. Sex differences in mental abilities: G masks the dimensions on which they lie. *Intelligence* **2007**, *35*, 23–39.
11. Ceci, S.J.; Williams, W.M.; Barnett, S.M. Women's underrepresentation in science: Sociocultural and biological considerations. *Psychol. Bull.* **2009**, *135*, 218–261.
12. Halpern, D.F.; Benbow, C.P.; Geary, D.C.; Gur, R.C.; Hyde, J.S.; Gernsbacher, M.A. The science of sex differences in science and mathematics. *Psychol. Sci. Public Interest* **2007**, *8*, 1–51.
13. Wai, J.; Lubinski, D.; Benbow, C.P. Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* **2009**, *101*, 817.
14. Keith, T.Z.; Reynolds, M.R.; Roberts, L.G.; Winter, A.L.; Austin, C. Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales—Second Edition. *Intelligence* **2011**, *39*, 389–404.
15. Lemos, G.C.; Abad, F.J.; Almeida, L.S.; Colom, R. Sex differences on g and non-g intellectual performance reveal potential sources of STEM discrepancies. *Intelligence* **2013**, *41*, 11–18.
16. Dykiert, D.; Gale, C.R.; Deary, I.J. Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence* **2009**, *37*, 42–47.
17. Brunner, M.; Krauss, S.; Kunter, M. Gender differences in mathematics: Does the story need to be rewritten? *Intelligence* **2008**, *36*, 403–421.
18. Feingold, A. Sex differences in variability in intellectual abilities: A new look at an old controversy. *Rev. Educ. Res.* **1992**, *62*, 61–84.
19. Irwing, P. Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personal. Individ. Differ.* **2012**, *53*, 126–131.

20. Keith, T.Z.; Reynolds, M.R.; Patel, P.G.; Ridley, K.P. Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence* **2008**, *36*, 502–525.
21. Lohman, D.F. *Cognitive Abilities Test*, Form 7; Riverside Publishing: Rolling Meadows, IL, USA, 2011.
22. Lohman, D.F.; Hagen, E. *Cognitive Abilities Test*, Form 6; Riverside Publishing: Itasca, IL, USA, 2001.
23. Lohman, D.F.; Lakin, J.M. Consistencies in sex differences on the Cognitive Abilities Test across countries, grades, test forms, and cohorts. *Br. J. Educ. Psychol.* **2009**, *79*, 389–407.
24. Lakin, J.M. Sex differences in reasoning abilities: Surprising evidence that male-female ratios in the tails of the quantitative reasoning distribution have increased. *Intelligence* **2013**, *41*, 263–274.
25. Carroll, J.B. *Human Cognitive Abilities: A Survey of the Factor-Analytic Studies*; Cambridge University Press: Cambridge, UK.
26. Lohman, D.F. *Cognitive Abilities Test*, Form 7. In *Research and Development Guide*; Riverside Publishing: Rolling Meadows, IL, USA, 2012.
27. Lohman, D.F.; Hagen, E. *Cognitive Abilities Test (Form 6) Research Handbook*; Riverside Publishing: Itasca, IL, USA, 2002.
28. Lohman, D.F.; Thorndike, R.L.; Hagen, E.; Smith, P.; Fernandes, C.; Strand, S. *Cognitive Abilities Test*; nferNelson: Windsor, UK, 2001.
29. Lakin, J.M.; Gambrell, J.L. Distinguishing verbal, quantitative, and nonverbal facets of fluid intelligence in young students. *Intelligence* **2012**, *40*, 560–570.
30. Lohman, D.F. The Wechsler Intelligence Scale for Children III and the Cognitive Abilities Test (Form 6): Are the General Factors the Same? (2003). Available online: http://faculty.education.uiowa.edu/dlohman (accessed on 9 January 2014).
31. Lohman, D.F. The Woodcock-Johnson III and the Cognitive Abilities Test (Form 6): A concurrent validity study. (2003). Available online: http://faculty.education.uiowa.edu/dlohman (accessed on 9 January 2014).
32. Deary, I.J.; Thorpe, G.; Wilson, V.; Starr, J.M.; Whalley, L.J. Population sex differences in IQ at age 11: The Scottish mental survey 1932. *Intelligence* **2003**, *31*, 533–542.
33. Hedges, L.V.; Friedman, L. Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Rev. Educ. Res.* **1993**, *63*, 94–105.
34. Wai, J.; Cacchio, M.; Putallaz, M.; Makel, M.C. Sex differences in the right tail of cognitive abilities: A 30 year examination. *Intelligence* **2010**, *38*, 412–423.
35. Halpern, D.F. Sex differences in intelligence. Implications for education. *Am. Psychol.* **1997**, *52*, 1091–102.
36. Lim, T.K. Gender-related differences in intelligence: Application of confirmatory factor analysis. *Intelligence* **1994**, *19*, 179–192.
37. It should be noted that even the paper folding task (Figure Analysis) on CogAT does not require spatial rotation because the items are designed to minimize the need to imagine rotation or hidden folds in the tasks. This is not true of all tests using this format.

38. Süβ, H.-M.; Beauducel, A. *Handbook of Understanding and Measuring Intelligence*; Wilhelm, O., Engle, R.W., Eds.; Thousand Oaks, CA, USA, 2005; pp. 314–332.
39. Lynn, R.; Irwing, P. Sex differences on the progressive matrices: A meta-analysis. *Intelligence* **2004**, *32*, 481–498.