CHILD DEVELOPMENT

Child Development, xxxx 2018, Volume 00, Number 0, Pages 1–14

More Similar Than Different: Gender Differences in Children's Basic Numerical Skills Are the Exception Not the Rule

Jane E. Hutchison, and Ian M. Lyons
University of Western Ontario and Georgetown University

Daniel Ansari
University of Western Ontario

This study investigates gender differences in basic numerical skills that are predictive of math achievement. Previous research in this area is inconsistent and has relied upon traditional hypothesis testing, which does not allow for assertive conclusions to be made regarding nonsignificant findings. This study is the first to compare male and female performance (N = 1,391; ages 6–13) on many basic numerical tasks using both Bayesian and frequentist analyses. The results provide strong evidence of gender similarities on the majority of basic numerical tasks measured, suggesting that a male advantage in foundational numerical skills is the exception rather than the rule.

The study of gender differences in math achievement has long been a topic of interest and has produced many influential, yet mixed, findings. In 1958, Anastasi concluded that boys begin to outperform girls in mathematics during the elementary school years. However, she also noted that gender differences may depend on contextual factors such as the type of mathematical problem being assessed. For example, girls were observed to excel on tasks of computational fluency, while boys were observed to excel on more cognitively demanding tasks such as problem solving. Early research replicated this pattern of results (Benbow & Stanley, 1980; Fennema & Carpenter, 1981; Marshall, 1984) and reviews of the literature reported consistent gender differences in math achievement (Fennema, 1974; Halpern, 1986; Maccoby & Jacklin, 1974). Consequently, researchers began to assess whether this consensus held when considering all of the evidence using meta-analytic approaches (Hyde, 2005; Hyde, Fennema, & Lamon, 1990). In the 1990s, an influential meta-analysis of 100 studies (3,175,188 participants) revealed virtually no gender differences in overall math performance and also demonstrated that the gender gap in math achievement had diminished over historical time (Hyde et al., 1990). More recent research suggests that this gap has continued to close, and that today, men and women display equal aptitude for mathematics (Hyde, 2005; Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Lindberg, Hyde, Petersen, & Linn, 2010).

Stereotypes About Gender Differences in Math Achievement

Although the gender gap in math performance has reportedly closed over time, gender gaps in math attitudes, perceptions, and interests remain prominent. More specifically, girls continue to report negative feelings concerning math (Nosek, Banaji, & Greenwald, 2002; Nosek & Smyth, 2011) and to perceive math as a "male subject" (Cvencek, Meltzoff, & Greenwald, 2011; Nosek et al., 2002). As a result, girls tend be less interested in quantitative fields such as Science, Technology, Engineering, and Mathematics (STEM) and therefore less likely to pursue a career in these areas (Eccles & Wang, 2016; Kiefer & Sekaquaptewa, 2007; Shapiro & Williams, 2012). It has been proposed that young children's attitudes, perceptions, and interests surrounding math are shaped early in development by environmental influences such as parent and teacher beliefs (Gunderson, Ramirez, Levine, & Beilock, 2012). One such belief includes the stereotype that boys are more likely to succeed in math than girls—a stereotype that

This research was supported by funding from the Canadian Institutes of Health Research (CIHR), The National Sciences and Engineering Research Council of Canada (NSERC), Canada Research Chairs Program (CRC) to Daniel Ansari, and a Banting Postdoctoral Fellowship Award to Ian M. Lyons.

Correspondence concerning this article should be addressed to Daniel Ansari, The University of Western Ontario, Westminster Hall, Room 325, London, Ontario N6A 3K7, Canada. Electronic mail may be sent to daniel.ansari@uwo.ca.

DOI: 10.1111/cdev.13044

continues to be held by parents and teachers alike (Cimpian, Lubienski, Timmer, Makowski, & Miller, 2016; Gunderson, Ramirez, Levine, et al., 2012; Jacobs, Davis-Kean, Bleeker, Eccles, & Malanchuk, 2005; Lavy & Sand, 2015; Riegle-Crumb & Humphries, 2012; Robinson-Cimpian, Lubienski, Ganley, & Copur-Gencturk, 2014). For example, Riegle-Crumb and Humphries (2012) found that teachers tend to perceive white boys as having greater mathematical abilities than their female counterparts and assumed that female students were performing worse in math than they actually were. Similarly, Cimpian et al. (2016) found that as early as first grade, teachers consistently underrate female, relative to male, math performance despite a general lack of evidence for differences in achievement across the genders (Cimpian et al., 2016). The authors interpreted these findings to suggest that teachers may be setting lower expectations for female math achievement leading to gender disparities in top-performing math students. In addition, these stereotypes continue to be perpetuated in the home. For example, Jacobs et al. (2005) observed that parents tend to provide more math-supportive environments for their sons compared with their daughters and that when fathers hold the stereotype that boys are better at math than grils, their daughter's interest in the subject decreases.

Therefore, empirical evidence suggests that in the 21st century, parents and teachers continue to hold the belief that boys are more likely to succeed in math compared with girls. Given that such beliefs have troubling consequences for female math education and career choices, it remains important to continue to conduct research that has the potential to address gender-related stereotypes in math.

The stereotype that boys are better at math than girls is most commonly reinforced by evidence indicating a persistent male advantage on standardized tests such as the mathematics section of the Scholastic Aptitude Test (SAT-M; Gallagher & Kaufman, 2005). However, as Spelke (2005) notes, it can be problematic to infer gender differences from such tests as they comprise many items that show varying gender disparities (Gallagher, Levin & Cahalan, 2002). Therefore, it is possible that items on tests of math achievement can be selected in such a way that may bias one gender over the other. Consequently, Spelke proposed that researchers should instead examine gender differences in more basic—both developmentally and in terms of cognitive complexity—numerical processing. This is especially germane as individual differences in these basic numerical processes have been shown to be predictive of math achievement in children and in adults (De Smedt, Noel, Gilmore, & Ansari, 2013; Feigenson, Libertus, & Halberda, 2013; Schneider et al., 2016). Moreover, as discussed earlier, gender stereotypes are prevalent among teachers and parents of young children. It is, therefore, critical to assess the degree to which gender differences exist in the basic number skills that young children are developing and that serve as part of the foundation upon which higher-level math skills are built. Ultimately, if there is any credence to the stereotype that boys are more likely to succeed in math than girls, one should expect to see a male advantage on the basic numerical tasks that are predictive of success in more advanced math. Against this background, the aim of the present study is to investigate if gender differences exist on commonly used measures of basic numerical processing.

Gender Differences in Basic Numerical Processing

Investigating gender differences in basic numerical processing is important for multiple reasons. For one, research in this area has expanded rapidly over the past three decades (De Smedt et al., 2013; Schneider et al., 2016). However, surprisingly little is known about how gender may influence the development of these early developing skills. Given that basic numerical processing is made up of many different components with different developmental trajectories and relationships to arithmetic achievement (Lyons, Price, Vaessen, Blomert, & Ansari, 2014), it is plausible that gender differences may manifest differently within various basic numerical skills. By systematically investigating gender differences across a series of widely used basic numerical tasks, we can provide a more refined picture about both gender similarities and differences in foundational numerical skills in the early years. This work could potentially inform how we might expect gender differences to play out in more complex numerical processing as each of the skills described earlier have been shown to be predictive of higher order math processing and are the focus of considerable developmental research trying to understand the foundational underpinnings of children's mathematical competencies (e.g., De Smedt et al., 2013; Feigenson et al., 2013; Schneider et al., 2016). Second, as previously discussed, the stereotype that boys are more likely to succeed in math continues to be widely held by both parents and teachers and such stereotypes have shown to be harmful for female math education and participation in STEM. Therefore, looking at gender differences within the foundational components of math can be informative for addressing the (potential) roots of gender differences in math and informing math-related gender stereotypes.

To date, research investigating gender differences in basic numerical processing has been surprisingly scarce. Of the studies that have examined gender differences in this area, very little difference has been observed with respect to overall performance on classic tasks of basic numerical processing (e.g., number comparison). Gender differences in number processing have been most consistently reported on measures that are thought to index the degree to which spatial processes influence numerical processing, such as the Spatial Numerical Association of Response Codes (SNARC) effect (i.e., one is quicker to process smaller numbers when responding with their left hand, and larger numbers when responding with their right hand) and number-line estimation (asking participants to estimate the relative spatial position of a numeral on a number line). The evidence in this area suggests that boys display a larger SNARC effect during both explicit and implicit number processing (Bull, Cleland, & Mitchell, 2013) and make more accurate and linear number estimations (Bull et al., 2013; Gunderson, Ramirez, Beilock, et al., 2012; Reinert, Huber, Nuerk, & Moeller, 2016; Thompson & Opfer, 2008). The male advantage in number-line estimation has been observed in both adults and children, and has been elicited by both bounded (i.e., the number line is marked with a start-, mid-, and end-point) and unbounded (i.e., the number line is marked only with a start-point and the length of one-unit) number-line tasks. Importantly, it should be noted that in the few studies that also collected measures of math achievement, this advantage did not translate into superior math performance on measures of numerical operations and math reasoning (Bull et al., 2013; Thompson & Opfer, 2008). Spatial processing is one of the few cognitive domains that shows reliable gender differences favoring boys (e.g., Halpern et al., 2007; Levine, Foley, Lourenco, Ehrlich, & Ratliff, 2016; Voyer, Voyer, & Bryden, 1995). One possibility, then, is that boys are more likely to rely on spatial approaches or strategies when completing basic numerical tasks.

In terms of more traditional tasks of basic numerical processing, the findings are much less consistent. For example, one study (*n* = 140) reported a gender difference favoring boys on a single- and multidigit number comparison task (Krinzinger, Wood, & Willmes, 2012), whereas another (*n* = 1,156) revealed a gender difference favoring girls on a similar task (Wei et al., 2012). In addition, in a sample of 526 seven- to sixteen-year olds, no gender differences were observed on a number comparison task and on three other basic numerical tasks, including reading numbers, writing numbers, and ordering numbers (Rosselli, Ardila, Matute, & Inozemtseva, 2009). In contrast, a different study (*n* = 220) observed a male advantage on a number writing task (Krinzinger, Kaufmann, et al., 2012). Gender differences have also been observed on the numerical distance effect (NDE), a measure derived from number comparison tasks wherein it is easier to discriminate between two numbers as the numerical distance between them increases (Bull et al., 2013). In a sample of 52 adult participants, the authors observed that boys were faster to discriminate between two numbers and that only girls displayed an NDE. Bull et al. (2013) interpreted this finding to suggest that boys have a more accurate mental representation of number, which makes it easier to discriminate between numbers that are closer together. However, there is now evidence against this interpretation of results from number comparison studies (Lyons, Nuerk, & Ansari, 2015). In addition, the lack of an NDE did not translate into a male advantage in terms of arithmetic performance, as the NDE did not correlate with performance on the arithmetic task.

The findings regarding gender differences in basic numerical processing may differ for multiple reasons. First, with the exception of Wei et al. (2012), the few studies that looked at gender differences within basic numerical skills comprised relatively small sample sizes, which are naturally more susceptible to greater variability in outcomes across studies. Second, many different tasks were used to assess what is sometimes thought of as the same basic number skill. For example, to assess number comparison, Rosselli et al. (2009) administered a task in which participants were presented with two cards, each containing eight three-digit numbers, and asked to state the largest number on each card. Comparatively, to assess presumably the same skill, Krinzinger, Wood, et al. (2012) administered both a single- and multidigit number comparison task in which children were asked to indicate which of two numbers was larger by pointing. Finally, Wei et al. (2012) administered a number comparison task in which participants were presented with pairs of single-digit numbers in varying sizes and asked to determine which was larger while ignoring the physical size of the number. The large degree of variety in the tasks used to assess gender differences in basic number skills could potentially explain the conflicting results in this area. Finally, of the studies that showed conflicting results, most focused on children between the ages of 7–10; although this does not explain the inconsistent findings, it does highlight the fact that none of the previous studies looked at gender differences across the elementary school years.

Taken together, the research concerning traditional basic numerical processing does not provide strong evidence for or against gender differences in this area. The lack of consistency across studies in terms of sample size and task implementation, in addition to the relatively limited scope in age groups studied, speaks to the need for a study that uses a large sample to systematically investigate gender differences in a wide range of basic numerical tasks across multiple developmental stages. This is precisely what we aim to do here.

Current Study

To sum, the majority of evidence investigating gender differences in math achievement converges around gender similarities; however, the same cannot be said for gender differences within the foundational components of more complex math processing. More specifically, of the relatively small handful of studies that have looked at gender differences in basic numerical skills, some have observed a gender difference favoring boys, others have observed an advantage favoring girls, and some have observed no difference at all. Overall, the fact that gender differences have been observed within some foundational components of math achievement suggests that gender differences in math may be more nuanced than current thinking would suggest. Put differently, not all basic numerical processing tasks index the same underlying processes, and thus, gender similarities and differences may vary as a function of the specific basic numerical processing task being examined. To this end, the present study aims to take a systematic approach to investigate gender differences on a series of frequently used basic numerical skills across multiple age groups.

If boys outperform girls on tasks that makeup the foundations of mathematical skills, this would lend support to the stereotype that boys have an early advantage that makes them more likely to succeed in math than girls. Conversely, if boys and girls perform equally on basic numerical tasks, it would suggest that both genders may be equally equipped (at least in terms of foundational numerical abilities) to acquire more complex math skills. Further, to help better contextual our results, we investigate how gender differences in basic numerical processing may differ as a function of age.

The Use of Bayesian Statistics to Assess Evidence for Gender Differences

Even in studies that have failed to detect gender differences in basic numerical processing (Aunio et al., 2004; Rosselli et al., 2009), the use of frequentist statistics does not allow for the conclusion to be made that the evidence favors the hypothesis that gender differences do not exist on these measures. More specifically, a statistically nonsignificant result could mean one of two things; that there is no effect or that there was not enough evidence to detect an effect. Traditional null-hypothesis testing does not afford differentiation between these two possibilities (Rouder, Speckman, Sun, Morey, & Iverson, 2009; van de Schoot et al., 2014; Wagenmakers, Morey, & Lee, 2016; Wagenmakers, Verhagen, & Ly, 2015). To rectify this issue, the present study takes a Bayesian approach to the study of gender differences in basic numerical processing. Bayesian statistics allows one to quantify evidence for the null hypothesis by weighing the evidence for the null against the evidence for the alternative hypothesis. The present study weighs the evidence for a gender difference on 12 basic numerical tasks (counting, dot comparison, dot estimation, number comparison, number-line estimation [100], number-line estimation [1,000], number ordering [one digit], number ordering [two digits], object matching, visual-audio matching, addition and subtraction, and multiplication and division), against the evidence for no gender difference. In sum, this approach allows us to state with greater confidence whether gender differences truly exist on these basic numerical tasks.

Method

It should be noted that the data presented here come from a large data set, a portion of which has been previously described and reported elsewhere (Bartelet, Ansari, Vaessen, & Blomert, 2014; Lyons & Ansari, 2015; Lyons et al., 2014, 2015). Importantly, while the overall data set remains the same, the present study addresses theoretical questions and employs statistical analyses distinct from those

Table 1
Brief Descriptions of the Tasks Administered

Task	Brief description
Number comparison (NumComp)	Determine which of two symbolic numbers is larger
Dot comparison (DotComp)	Determine which of two dot arrays contains more
Number ordering (NumOrd)	Determine whether three symbolic numbers presented horizontally are in the correct ascending order. Single- and double-digit sequences were administered separately
Counting (Counting)	Count number of dots (1–9) on the screen as quickly and as accurately as possible
Dot estimation (DotEst)	Estimate how many dots makeup a visually presented array of dots
Number-line estimation (NumLine)	Indicate on a horizontal line (either 0–100 or 0–1000) where a symbolic number should fall
Addition/subtraction (AddSub)	Addition and subtraction subtasks of a standardized test of arithmetic ability
Multiplication/division (MultDiv)	Multiplication and division subtasks of a standardized test of arithmetic ability
Object matching (ObjMatch)	Determine which of two sets of common objects contain the same number of objects as a sample array
Visual-audio matching (VisAud)	Determine if a number spoken out loud matches a visually presented number
Stimulus–Response processing (StimResp)	Press the one marked square out of four as quickly as possible
Reading (Reading)	Standardized measure of reading ability
Ravens (Ravens)	Standardized assessment of nonverbal (spatial) reasoning/IQ

Note. Task abbreviations in parentheses are used throughout the remainder of the manuscript.

studies cited earlier. For an in-depth description of the procedure and tasks, see Lyons et al. (2014). Brief descriptions of the tasks are displayed in Table 1.

Participants

Data were collected from 1,463 children in Grades 1–6 from seven different primary schools in the Netherlands. Parents either provided or withheld consent by returning the appropriate form. Children who performed at chance on any of the tasks for which chance can be defined (> 49% error rate on any of the binary forced-choice numerical tasks: NumComp, DotComp, NumOrd, VisAud, ObjMatch; > 24% error rate on the four-choice StimResp task) were removed from further analysis, as chance performance is difficult to interpret. This led to the removal of 37 children (2.53%) from further analyses. For each grade, outliers were removed by checking whether a child's score on a given task was more or < 4 *SD*s from the mean for that task. This led to the removal of 35 additional children from the analysis (2.39%). A total of 72 children were ultimately removed from the data set resulting in a final sample size of $N = 1{,}391$ (722 female; see Table S1 for a breakdown of *n*s by grade). The data collection was approved by the Ethics Review Board at Maastricht University.

Task Scoring

On the majority of the tasks administered (NumComp, DotComp, NumOrd [one digit and two digits], VisAud, ObjMatch, Counting and StimResp), performance scores were calculated as a composite of error rates and response times (correct trials only), using the formula: $P = \mathrm{RT}(1 + 2\mathrm{ER})$, where a higher value indicates worse performance. Error rates were multiplied by 2 because most tasks were constrained by binary responses (ER = 0.5 indicates chance). Combining response times and error rates provides a more complete picture of overall performance, as it reduces the number of statistical tests by half, which in effect reduces the risk of false positives and also controls for variability in speed-accuracy trade-offs across tasks. Essentially, the combined score is a measure of reaction times (ms) after they have been penalized for inaccurate performance. The scale ranges between a participant's actual average response time (where $P = \mathrm{RT}$) for completely accurate performance (0% errors) and twice that value ($P = 2\mathrm{RT}$) for chance performance (50% errors).

Performance was calculated differently for the DotEst and NumLine (NL100 and NL1000) tasks, as these tasks include many trials on which exactly correct answers are unlikely, which makes traditional error rates difficult to interpret. Performance for these tasks was, therefore, calculated using percent absolute errors: $\mathrm{PAE} = |\mathrm{Est} - \mathrm{Target}|/\mathrm{Scale}$, where Est is the child's estimation, Target is the target number, and Scale is the scale or range of target numbers. Scale was 100 for NL100, 1000 for NL1000, and 16 for DotEst. Final scores were calculated by averaging the PAE for each trial, with a higher number indicating worse performance.

Analyses

The aim of the present study was to investigate gender differences in performance on basic numerical tasks. We did so using both traditional hypothesis testing and Bayesian analyses. Following traditional frequentist methods, we ran a series of univariate analyses of variance (ANOVAs) in SPSS (IBM Corp, 2013) with task as the dependent variable and gender (0 = *male*, 1 = *female*) and grade as fixed factors. To control for multiple comparisons, we used the Dunn-Šidak (Šidák, 1967) corrected significance threshold of $p = .003$ (this value was derived using the following formula: $1 - (1 - \alpha)^{1/k}$, where $\alpha = .05$ and $k = 15$ (k refers to the number of independent significance tests that were run). To examine how gender differences may differ depending on grade, we ran a series of t tests comparing performance as a function of gender for each task per grade (corrected significance threshold: $p = .0006$). Using traditional hypothesis testing, we can infer from a significant result that a gender difference may exist on a given task; however, we cannot infer from a nonsignificant result that there is evidence in support of the null hypothesis. We, therefore, furthered our investigation of gender differences in basic numerical processing through Bayesian analyses.

Bayesian analyses weigh the evidence for the alternative (B_{10}; evidence for the existence of a gender difference), against the evidence for the null (B_{01}; evidence for the lack of a gender difference). The resulting statistics provide an index of the strength of the evidence for or against the alternative (i.e., the Bayes factor, BF). This allows us to infer how likely it is that a gender difference truly exists on any of our 15 measures. We ran Bayesian ANOVAs for each task with task performance as the dependent variable and gender and grade as fixed factors using the software JASP (JASP Team, 2016). This provided us with B_{10} and B_{01} factors for the main effect of gender and the Gender × Grade interaction. To quantify the evidence for the alternate hypothesis (that gender differences do exist), we used the default Cauchy distribution prior centered on the null with a width of 0.707. Given the dearth of prior evidence relevant to the present investigation of gender differences in basic number processing, we chose this default prior rather than an informative prior. To quantify evidence for the null hypothesis (that there are no gender differences), we set the prior as an effect size of 0 (which again is the default in JASP for calculating BF01).

Results

Frequentist Results

Table 2 displays the results from the multiple univariate ANOVAs. A significant main effect of grade was observed for each task, which simply indicated that older children performed better on each task (see Table S1 for mean task performance). In addition, a significant main effect of gender favoring boys was observed on the NumOrd (two digits), NL100, NL1000, and AddSub tasks. Furthermore, an interaction between gender and grade was observed on the Counting, NL100, and NL1000 tasks, indicating that gender differences varied by grade for these tasks. A series of t tests (displayed in Table 3) revealed a significant gender difference favoring girls in Counting in Grade 1 but none of

Table 2
Results From the Univariate Analyses of Variance Investigating the Effects of Gender and Grade on Task Performance

		Gender	Grade	Interaction
NumComp	*F*	7.05	**318.14**	0.23
	p	.008	**<.001**	.950
DotComp	*F*	1.67	**121.81**	2.44
	p	.196	**<.001**	.033
NumOrd (one digit)	*F*	1.50	**243.40**	0.72
	p	.220	**<.001**	.612
NumOrd (two digits)	*F*	**10.81**	**133.93**	0.76
	p	**.001**	**<.001**	.554
Counting	*F*	8.49	**258.78**	**5.92**
	p	.004	**<.001**	**<.001**
DotEst	*F*	2.11	**73.98**	0.31
	p	.147	**<.001**	.910
NumLine (0–100)	*F*	**53.58**	**448.96**	**8.79**
	p	**<.001**	**<.001**	**<.001**
NumLine (0–1,000)	*F*	**118.39**	**143.84**	**8.39**
	p	**<.001**	**<.001**	**<.001**
AddSub	*F*	**28.55**	**550.30**	1.10
	p	**<.001**	**<.001**	.357
MultDiv	*F*	1.84	**62.36**	0.63
	p	.176	**<.001**	.596
ObjMatch	*F*	3.93	**278.58**	2.02
	p	.048	**<.001**	.073
VisAud	*F*	2.39	**447.91**	0.378
	p	.122	**<.001**	.864
StimResp	*F*	0.66	**285.74**	1.41
	p	.418	**<.001**	.219
Reading	*F*	0	**447.82**	0.82
	p	.970	**<.001**	.532
Ravens	*F*	0	**77.37**	0.21
	p	.992	**<.001**	.957

Note. The Bonferonni corrected significance threshold was $p < .003$. Significant results are bolded.

the other grades; a significant gender difference favoring boys in the NL100 task in Grades 1 and 2, but none of the other grades; and a significant gender difference favoring boys in the NL1000 task in Grades 2–5, but not in Grade 6.

Although it is informative to know on which tasks and in which grades gender has a significant effect, it is also crucial to assess the strength of such effects. Effect sizes reflecting the magnitude of overall and grade-specific gender differences in task performance are displayed in Table 4. As is reflected in the table, the overall gender difference favoring boys in the NumOrd (two digits) and AddSub tasks appear to be relatively small. In addition, when broken down by grade, the effect of gender on NumOrd (two digits) and AddSub performance is no longer significant, suggesting that within each grade, boys and girls are performing equally on both measures. Gender appears to have the strongest effect on overall performance in the NumLine tasks. When looking across the grades, the male advantage in the NL100 task appears to strengthen between Grades 1 and 2, but it drops below significance in Grades 3–6. The gender difference favoring boys in the NL1000 task appears to be the strongest and most consistent; however, the strength of the effect appears to decrease as grade increases, and by Grade 6, the effect of gender is no longer significant. To further assess the magnitude of the significant gender differences, and to quantify evidence in favor of the null, especially in the case of nonsignificant results, we followed up the traditional hypothesis testing with Bayesian analyses.

Table 3

Results From the t *Tests Investigating the Difference in Performance Between Boys and Girls on Each Task Within Each Grade*

		Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
NumComp	*t*	−0.34	−0.55	−1.52	−1.70	−2.92	−1.36
	p	.734	.584	.130	.090	.004	.176
DotComp	*t*	2.21	1.99	−0.73	−1.05	−0.30	0.29
	p	.028	.047	.464	.294	.762	.776
NumOrd (one digit)	*t*	−0.65	0.73	−1.67	0.19	−1.18	−1.17
	p	.516	.464	.096	.849	.240	.245
NumOrd (two digits)	*t*	—	−1.36	−2.31	−0.272	−2.13	1.38
	p	—	.177	.021	.785	.035	.170
Counting	*t*	**3.80**	1.71	0.39	−0.28	−1.26	0.06
	p	**< .001**	.088	.700	.776	.207	.951
DotEst	*t*	0.10	−1.16	−1.08	−0.76	−0.15	−0.63
	p	.920	.248	.280	.448	.882	.529
NumLine (0–100)	*t*	**−3.97**	**−4.36**	−1.83	−2.22	−2.44	−2.10
	p	**< .001**	**< .001**	.068	.027	.015	.037
NumLine (0–1,000)	*t*	—	**−5.83**	**−6.06**	**−4.85**	**−4.13**	−1.74
	p	—	**< .001**	**< .001**	**< .001**	**< .001**	.083
AddSub	*t*	−0.43	−1.42	−3.03	−2.42	−2.936	−2.62
	p	0.67	.159	.003	.016	.004	.009
MultDiv	*t*	—	—	0.46	−1.03	−0.68	−1.38
	p	—	—	.644	.305	.497	.170
ObjMatch	*t*	1.46	2.23	−0.31	0.69	1.27	0.57
	p	.147	.027	.755	.493	.206	.568
VisAud	*t*	−0.05	−1.11	−0.23	−0.01	−1.83	−1.25
	p	.961	.267	.816	.989	.069	.213
StimResp	*t*	1.63	−0.65	−0.83	−1.00	−0.63	−1.51
	p	.104	.515	.408	.316	.531	.132
Reading	*t*	0.89	0.843	−0.17	0.04	−1.35	−0.64
	p	.375	.400	.867	.968	.177	.523
Ravens	*t*	0.44	0.316	−0.75	−0.25	0.07	0.05
	p	.661	.752	.453	.804	.944	.958

Note. The Bonferonni corrected significance threshold was $p < .0006$. Significant results are bolded. Because lower scores reflect better performance in all tasks except for AddSub, MultDiv, Reading, and Ravens, the signs for these four tasks were switched so that for all tasks a negative value reflects better male performance, and a positive value reflects better female performance.

Table 4
Effect Sizes (Cohen's ds*) Reflecting the Magnitude of the Difference Between Male and Female Performance on Each Task*

	All	Grade × Gender (*p*)	Grade1	Grade2	Grade3	Grade4	Grade5	Grade6
NumComp	−0.14	.950	−0.05	−0.08	−0.19	−0.22	−0.38	−0.18
DotComp	0.07	.033	0.31	0.28	−0.09	−0.13	−0.04	0.04
NumOrd (one digit)	0.07	.612	−0.09	0.10	−0.21	0.02	−0.15	−0.15
NumOrd (two digits)	**−0.18**	.554		−0.19	−0.30	−0.03	−0.27	−0.20
Counting	0.16	**< .001**	**0.54**	0.26	0.05	−0.04	−0.16	0.01
DotEst	−0.08	.910	0.01	−0.16	−0.14	−0.10	−0.02	−0.08
NumLine (0–100)	**−0.39**	**< .001**	**−0.55**	**−0.63**	−0.23	−0.28	−0.32	−0.27
NumLine (0–1,000)	**−0.59**	**< .001**		**−0.89**	**−0.80**	**−0.69**	**−0.59**	−0.23
AddSub	**−0.29**	.357	−0.06	−0.20	−0.38	−0.31	−0.38	−0.34
MultDiv	−0.07	.596			0.06	−0.13	−0.09	−0.18
ObjMatch	0.11	.073	0.20	0.33	−0.04	0.09	−0.16	0.07
VisAud	0.08	.864	−0.01	−0.16	−0.03	0.00	−0.24	−0.16
StimResp	0.04	.219	0.23	−0.09	−0.10	−0.13	−0.08	−0.20
Reading	0.00	.532	0.12	0.12	−0.02	0.01	−0.18	−0.08
Ravens	0.00	.957	0.06	0.05	−0.09	−0.03	0.01	0.01

Note. The effect sizes in the "all" column were calculated from *F*-statistics (shown in Table 2), whereas the effect sizes in the grade columns were calculated from *t*-statistics (shown in Table 3). Significant values are bolded. Because lower scores reflect better performance in all tasks except for AddSub, MultDiv, Reading, and Ravens, the signs for these four tasks were switched so that on all tasks a positive effect size (the color blue) indicates better performance favoring girls, whereas a negative effect size (the color red) indicates better performance favoring boys (see online version for color version of the table).

Bayesian Results

Figure 1 displays the BFs for the main effect of gender (see Table S2 for a list of exact values and BFs for the interaction term). The B_{10} values reflect the strength of the evidence for the alternative (a gender difference does exist), whereas the B_{01} values reflect the strength of the evidence for the null (a gender difference does not exist). According to Jeffreys (1961), the interpretation of BFs can be organized into three categories: if less than 3, the evidence is anecdotal (i.e., the evidence is not sufficient); if above 3, the evidence is substantial; and if above 10, the evidence is strong. As can be seen in Figure 1, there is negligible evidence for both the alternative and the null on the AddSub, NumOrd (two digits), and ObjMatch tasks. Therefore, we do not have sufficient evidence to infer whether a gender difference exists on these measures. However, on the remaining 12 tasks, the evidence clearly supports either the alternative or the null. First, on the NumComp, DotComp, NumOrd (one digit), DotEst, MultDiv, VisAud, StimResp, Reading, and Ravens tasks, there is substantial to very strong evidence in support of the null hypothesis and very weak evidence for the alternative, meaning that it is very unlikely that gender influences performance on these measures. On the remaining three tasks (NL100, NL1000, and Counting), there was substantial evidence for an effect of gender on the NL100 and Counting tasks, and very strong evidence for a gender difference in performance on the NL1000 task. However, in all cases in which the evidence supports the alternative, there is very strong evidence for a Grade × Gender interaction, suggesting that the gender differences observed on these three measures are not consistent across the grades (Table S2).

Discussion

Despite evidence indicating gender similarities in math, the stereotype that boys are more likely to succeed in math than girls continues to be widely held. Such stereotypes, when held by parents, teachers, and students themselves, can be harmful for female math education and may ultimately discourage girls from pursuing careers in STEM fields. If this stereotype is in fact true, a male advantage should be observed in the basic numerical skills that makeup the foundation of more advanced math skills. To date, research in this area has been scarce and the findings that have been reported are inconsistent. Moreover, the majority of previous work has employed primarily traditional hypothesis testing to assess gender differences on basic numerical tasks; however, traditional hypothesis testing

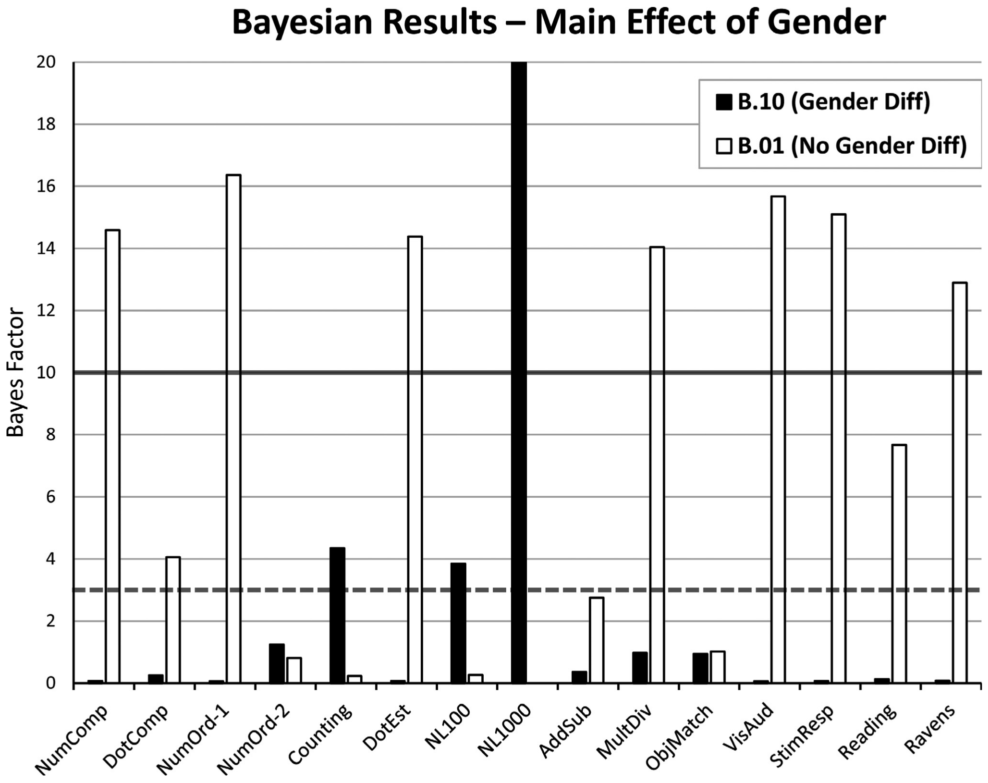

Figure 1. Bayesian results—main effect of gender.
Notes. A Bayes factor (BF) < 3 (dashed gray line) is considered anecdotal; 3 < BF < 10 is considered substantial; BF > 10 (thick gray line) is considered strong. For exact values and BF values for the interaction term, see Table S2.

does not allow for assertive conclusions to be made concerning nonsignificant findings (e.g., it cannot be used to quantify the evidence in support of the null hypothesis). The present study is, to our knowledge, the first to compare male and female performance in basic numerical processing using a single, large sample of children (1,391) from both frequentist and Bayesian perspectives.

Overall, the evidence from the present study largely supports the null hypothesis (i.e., a gender difference does not exist) for the majority of tasks measured here. In other words, the presence of gender differences in basic numerical processing in children appears to be more the *exception* than the rule. Frequentist analyses revealed a nonsignificant difference between male and female performance on the majority of the tasks administered (numerical comparison, dot comparison, numerical ordering [one and two digits], dot estimation, multiplication/division, object matching, visual-audio matching, stimulus–response processing, reading ability, and nonverbal intelligence). The nonsignificant findings were further quantified through Bayesian analyses, which affirmed that the evidence for the null was substantial to very strong and that the evidence for the alternative was very weak. It is, therefore, highly unlikely that gender influences performance on any of the aforementioned tasks in children in Grades 1–6. In contrast, the frequentist analyses revealed a significant gender difference on the numerical ordering (two digits) and addition/subtraction tasks. However, the associated effect sizes were relatively weak and when broken down by grade, boys and girls appeared to perform equally well on both of these measures. Moreover, Bayesian analyses showed that the evidence for a gender difference on these tasks was insufficient (e.g., there is not enough evidence in support of either the null or the alternate hypothesis). That is to say, in the present study, there was not enough evidence to determine whether gender influences performance on the numerical ordering (two digits) or addition/subtraction tasks, and therefore, we cannot draw firm conclusions regarding the effect of gender on either of these measures. We find evidence for gender differences on only three of the basic numerical tasks assessed here: the two number-line tasks (0–100 and 0–1,000) and the counting task, and only the two number-line tasks showed an advantage for boys over girls. Furthermore, these effects *decreased* with development (i.e., gender differences were smaller in older relative to younger children). Taken together, the present findings indicate that, for the most part, gender does not influence basic numerical processing.

With the exception of number-line estimation, prior research looking at gender differences in basic numerical skills has been somewhat mixed. Therefore, it is unclear how our findings regarding more traditional tasks of basic numerical processing (e.g., number comparison) fit in with the current literature. For example, our null findings converge with those reported by Rosselli et al. (2009) who observed nonsignificant gender differences on both a number comparison and number ordering task. However, these findings diverge from what was reported by Krinzinger, Wood, et al. (2012), who observed a male advantage in number comparison, and Wei et al. (2012) who observed a female advantage on a similar task. As noted earlier, most of these studies comprise relatively small sample sizes with limited age ranges and utilize a wide variety of tasks to assess ostensibly the same basic numerical skills; hence, it is difficult to pinpoint why previous findings differ from ours.

In terms of the tasks in which we did observe a gender difference (i.e., number-line estimation and counting), the findings concerning number-line estimation converge with past studies suggesting that number-line estimation is one of the few numerical tasks that does show an effect of gender favoring boys (Bull et al., 2013; Gunderson, Ramirez, Beilock, et al., 2012; Reinert et al., 2016; Thompson & Opfer, 2008). However, the current results are the first to suggest that the male advantage in number-line estimation may decrease with age. In terms of counting, to our knowledge, the current findings are the first to indicate a female advantage on this task specifically within the first grade. We further expand upon the role of gender in number-line estimation and counting in the following paragraphs.

In terms of number-line estimation, Bayesian analyses revealed substantial evidence for a gender difference on the NL100 task and very strong evidence for a gender difference on the NL1000 task. The gender difference observed in number-line estimation is potentially related to a male advantage in visual-spatial skills (for a review on gender differences in cognition, see Halpern et al., 2007), as such skills have previously been linked to performance on number-line tasks (Gunderson, Ramirez, Beilock, et al., 2012). However, it should be noted that the effect of gender in number-line performance decreased with grade. More specifically, the male advantage in number-line estimation was no longer significant by Grade 4 for the NL100 task and by Grade 6 for the NL1000 task.

Although a male advantage on the number-line task is not all too surprising, the fact that this advantage decreased with grade is of potential interest. We offer two speculative interpretations. For one, it is possible that cultural factors such as education are helping to mitigate against gender differences in visual-spatial processing, allowing for girls to catch up with their male counterparts on the number-line estimation task in later grades.

Alternatively, it could be that when completing the number-line estimation task, older children rely less heavily on spatial strategies—which are prone to gender differences—to complete the task. A recent study investigating adult gender differences in number-line estimation offers some support for this interpretation (Reinert et al., 2016). In this study, the authors did not observe a gender difference on a traditional number-line estimation task, which converges with what we observed in older children. However, when the authors increased the visual-spatial demands of the task by presenting participants with an unbounded number line (only a start-point and the length of one-unit are indicated), a male advantage was observed. This suggests that the visual-spatial demands elicited by the traditional number-line estimation task may not be strong enough to result in a male advantage for adults and perhaps even older children. Future research should further investigate the mediating role of visual-spatial processing in number-line estimation and how it may change with age.

The third instance in which Bayesian analyses revealed substantial evidence for a gender difference was on the counting task, in which girls significantly outperformed boys in counting in Grade 1 but none of the other grades. The female advantage in first-grade counting could potentially be explained by the fact that girls tend to have an advantage in verbal abilities (Halpern et al., 2007). However, this interpretation is limited by the fact that we do not see greater female performance on the reading task, which we would expect if the girls in our sample had a verbal advantage. In addition, we only see a gender difference in counting in Grade 1; if an advantage in verbal fluency contributed to the effect of gender on this task, we might expect the female advantage in counting to be more consistent. It is, therefore, difficult to interpret why we see a female advantage for counting in Grade 1; nevertheless, such a finding does not align with the stereotype that boys are more likely to succeed in math than girls.

Overall, the current findings point toward gender similarities on the majority of tasks administered. Prior research investigating gender differences in basic numerical processing has relied solely upon frequentist statistical testing methods, which does not allow for assertive conclusions to be made regarding nonsignificant findings. Scientists who are perhaps all too aware of this crucial but often overlooked shortcoming of frequentist approaches will (rightfully) tend to be cautious when characterizing null findings to members of the media and general public, which may contribute—inadvertently—to the persistence of gender stereotypes in math. Therefore, when examining gender differences, it is both highly informative and perhaps even conceptually transformative to be able to quantify evidence for the null (i.e., gender similarities). To this end, our study is the first to provide compelling evidence *for* gender similarities across a range of elementary ages and wide range of basic numerical skills.

It is especially important to provide strong, clear evidence in favor of gender equality in children, because, as previously mentioned in the Introduction, teachers and parents continue to hold the belief that girls have poorer math skills, and such stereotypes are known to be damaging for female math education and career choices (Cvencek et al., 2011; Eccles & Wang, 2016; Good, Aronson, & Harder, 2008; Kiefer & Sekaquaptewa, 2007; Nosek & Smyth, 2011; Nosek et al., 2002; Shapiro & Williams, 2012; Tomasetto, Alparone, & Cadinu, 2011). The current findings help to dispel this stereotype by suggesting that, at the level of basic numerical processing, both boys and girls should be equally capable of acquiring more complex math skills. Ultimately, while certain stereotypes may have some basis in fact, many are predicated largely on myth. By examining gender differences across a wide range of basic numerical tasks, we are able to clearly state that gender differences are the exception not the rule.

Limitations

It should be noted that the findings from the present study are specific to children growing up in the Netherlands. As previously mentioned, cross-national research has shown that gender differences in math vary across cultures, with some countries displaying gender similarities and others displaying gender differences (Else-Quest, Hyde, & Linn, 2010). In the country in which the data were collected, there does not appear to be strong evidence for a male advantage in math, although gender differences in math have been observed in other countries. Thus, it is plausible that we may see greater gender differences in basic numerical processing in countries that report stronger gender differences in math. Therefore, to culturally situate the current findings, an important future direction is to investigate gender differences in basic numerical skills in regions in which stronger gender differences in math have been reported.

In addition, due to the difficulty of the tasks, the findings of the present study are specific to children who have already begun formal schooling. An important future direction would be to investigate gender differences in (age-appropriate) basic numerical skills of younger children, as gender may differentially affect the relevant skills of children who are just beginning to acquire them.

Finally, given the fact that basic numerical skills makeup the foundation upon which more complex math abilities develop, we have interpreted the lack of gender differences in these skills to suggest that boys and girls should be equally capable of acquiring more complex math abilities. However, it is important to note that basic numerical skills do not account for all of the variation in later math achievement. Therefore, it is important to understand how gender may be acting through other cognitive correlates of complex math processing to influence math performance in order to further address gender stereotypes in math.

Conclusion

In conclusion, the present study is the first to investigate gender differences in basic numerical processing in a large sample of children using both frequentist and Bayesian analyses. Moreover, our findings are the first to provide strong evidence against the effect of gender on many basic numerical tasks. Of the three instances in which strong gender differences were observed (number line 0–100, number line 0–1,000, and counting), only the number-line tasks favored boys, and in each case, the effect of gender decreased with grade, becoming nonsignificant in older children. In sum, the presence of gender differences in basic numerical processing in children is more the exception than the rule. The strong evidence for gender similarities in basic numerical processing and the underwhelming support for the alternative suggest that boys and girls are equally equipped with basic numerical competencies and thus should be equally capable of acquiring complex mathematical skills. Such findings may have the potential to dissuade parents and teachers from underestimating the capacity for girls to excel in math.

References

Anastasi, A. (1958). *Differential psychology* (3rd ed.). New York, NY: Macmillan.

Aunio, P., Ee, J., Lim, S. E. A., Hautamäki, J., Luit, V., & Johannes, E. H. (2004). Young children's number sense in Finland, Hong Kong and Singapore. *International Journal of Early Years Education*, *12*, 196–216. https://doi.org/10.1080/0966976042000268681

Bartelet, D., Ansari, D., Vaessen, A., & Blomert, L. (2014). Cognitive subtypes of mathematics learning difficulties in primary education. *Research in Developmental Disabilities*, *35*, 657–670. https://doi.org/10.1016/j.ridd.2013.12.010

Benbow, C. P., & Stanley, J. C. (1980). Gender differences in mathematical ability: Fact or artifact? *Science*, *210*, 1262–1264. https://doi.org/10.1126/science.7434028

Bull, R., Cleland, A. A., & Mitchell, T. (2013). Gender differences in the spatial representation of number. *Journal of Experimental Psychology: General*, *142*, 181–192. https://doi.org/10.1037/a0028387

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K Cohorts. *AERA Open*, *2*, https://doi.org/10.1177/2332858415616358

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*, 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

De Smedt, B., Noel, M. P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*, 48–55. https://doi.org/10.1016/j.tine.2013.06.001

Eccles, J. S., & Wang, M. T. (2016). What motivates females and males to pursue careers in mathematics and science? *International Journal of Behavioral Development*, *40*, 100–106. https://doi.org/10.1177/0165025415616201

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103–127. https://doi.org/10.1037/a0018053

Feigenson, L., Libertus, M. E., & Halberda, J. (2013). Links between the intuitive sense of number and formal mathematics ability. *Child Development Perspectives*, *7*, 74–79. https://doi.org/10.1111/cdep.12019

Fennema, E. (1974). Mathematics learning and the genderes. *Journal for Research in Mathematics Education*, *5*, 126–129. https://doi.org/10.2307/748949

Fennema, E., & Carpenter, T. P. (1981). Gender-related differences in mathematics: Results from the National Assessment. *Mathematics Teacher*, *74*, 554–559. https://doi.org/10.2307/748949

Gallagher, A., Levin, J., & Cahalan, C. (2002). Cognitive patterns of gender differences on mathematics admissions tests. *ETS Research Report Series*, *2002*(2). https://doi.org/10.1002/j.2333-8504.2002.tb01886.x

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, *29*, 17–28. https://doi.org/10.1016/j.appdev.2007.10.004

Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: The role of the linear number-line. *Developmental Psychology*, *48*, 1229. https://doi.org/10.1037/a0027433

Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Gender Roles*, *66*, 153–166. https://doi.org/10.1007/s11199-011-9996-2

Halpern, D. F. (1986). *Gender differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of gender differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581. https://doi.org/10.1037/0003-066x.60.6.581

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139. https://doi.org/10.1037/0003-066x.60.6.581

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*, 494–495. https://doi.org/10.1126/science.1160364

IBM Corp. (2013). *IBM SPSS statistics for windows, version 22.0*. Armonk, NY: IBM Corp.

Jacobs, J. E., Davis-Kean, P., Bleeker, M., Eccles, J. S., & Malanchuk, O. (2005). I can, but I don't want to: The impact of parents, interests, and activities on gender differences in math. In A. Gallagher & J. Kaufman (Eds.), *Gender difference in mathematics* (pp. 246–263). New York, NY: Cambridge University Press.

JASP Team. (2016). *JASP (Version 0.7.5.5)* [Computer software]. Amsterdam, The Netherlands..

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York, NY: Oxford University Press.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes a prospective study of female college students. *Psychological Science*, *18*, 13–18. https://doi.org/10.1111/j.1467-9280.2007.01841.x

Krinzinger, H., Kaufmann, L., Gregoire, J., Desoete, A., Nuerk, H. C., & Willmes, K. (2012). Gender differences in the development of numerical skills in four European countries. *International Journal of Gender, Science and Technology*, *4*, 62–77. Retrieved from: http://genderandset.open.ac.uk/index.php/genderandset/article/view/155

Krinzinger, H., Wood, G., & Willmes, K. (2012). What accounts for individual and gender differences in the multi-digit number processing of primary school children? *Zeitschrift für Psychologie*. https://doi.org/10.1027/2151-2604/a000099

Lavy, V., & Sand, E. (2015). *On the origins of gender human capital gaps: Short and long- term consequences of teachers' stereotypical biases* (No. w20909). National Bureau of Economic Research. https://doi.org/10.3386/w20909

Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., & Ratliff, K. (2016). Gender differences in spatial cognition: Advancing the conversation. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*, 127–155. https://doi.org/10.1002/wcs.1380

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*, 1123. https://doi.org/10.1037/a0021276

Lyons, I. M., & Ansari, D. (2015). Numerical order processing in children: From reversing the distance-effect to predicting arithmetic. *Mind, Brain, and Education*, *9*, 207–221. https://doi.org/10.1111/mbe.12094

Lyons, I. M., Nuerk, H., & Ansari, D. (2015). Rethinking the implications of numerical ratio effects for understanding the development of representational precision and numerical processing across formats. *Journal of Experimental Psychology: General*, *144*, 1021–1035. https://doi.org/10.1037/xge0000094

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental Science*, *17*, 714–726. https://doi.org/10.1111/desc.12152

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of gender differences*. Stanford, CA: Stanford University Press.

Marshall, S. P. (1984). Gender differences in children's mathematics achievement: Solving computations and story problems. *Journal of Educational Psychology*, *76*, 194. https://doi.org/10.1037/0022-0663.76.2.194

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44. https://doi.org/10.1037/0022-3514.83.1.44

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal*, *48*, 1125–1156. https://doi.org/10.3102/0002831211410683

Reinert, R. M., Huber, S., Nuerk, H. C., & Moeller, K. (2016). Gender differences in number-line estimation: The role of numerical estimation. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12203

Riegle-Crumb, C., & Humphries, M. (2012). Exploring bias in math teachers' perceptions of students' ability by gender and race/ethnicity. *Gender & Society*, *26*, 290–322. https://doi.org/10.1177/0891243211434614

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, *50*, 1262. https://doi.org/10.1037/a0035073

Rosselli, M., Ardila, A., Matute, E., & Inozemtseva, O. (2009). Gender differences and cognitive correlates of mathematical skills in school-aged children. *Child Neuropsychology*, *15*, 216–231. https://doi.org/10.1080/09297040802195205

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. https://doi.org/10.3758/pbr.16.2.225

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2016). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*. https://doi.org/10.1111/desc.12372

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Gender Roles*, *66*, 175–183. https://doi.org/10.1007/s11199-011-0051-0

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626–633. https://doi.org/10.1080/01621459.1967.10482935

Spelke, E. S. (2005). Gender differences in intrinsic aptitude for mathematics and science: A critical review. *American Psychologist*, *60*, 950. https://doi.org/10.1037/0003-066x.60.9.950

Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and gender differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology*, *101*, 20–51. https://doi.org/doi: 10.1016/j.jecp.2008.02.003

Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls' math performance under stereotype threat: The moderating role of mothers' gender stereotypes. *Developmental Psychology*, *47*, 943–949. https://doi.org/10.1037/a0024047

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken Marcel, A. G. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, *85*, 842–860. https://doi.org/10.1111/cdev.12169

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of gender differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250. https://doi.org/10.1037/0033-2909.117.2.250

Wagenmakers, E., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current*

Directions in Psychological Science, *25*, 169–176. https://doi.org/10.1177/0963721416643289

Wagenmakers, E. J., Verhagen, J., & Ly, A. (2015). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 1–14. https://doi.org/10.3758/s13428-015-0593-0

Wei, W., Lu, H., Zhao, H., Chen, C., Dong, Q., & Zhou, X. (2012). Gender differences in children's arithmetic performance are accounted for by gender differences in language abilities. *Psychological Science*, *23*, 320–330. https://doi.org/10.1177/0956797611427168

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Table S1. Mean Task Performance for Boys and Girls

Table S2. Bayes Factors for the Main Effect of Gender and the Gender × Grade Interaction