Boy's math performance, compared to girls', jumps at age 6 (in the ELFE's data at least)

Article *in* British Journal of Developmental Psychology · June 2022

DOI: 10.1111/bjdp.12423

2 authors, including:

Jean-Paul Fischer
University of Lorraine

Boy's math performance, compared to girls', jumps at age 6 (in the ELFE's data at least)

Jean-Paul Fischer, Xavier Thierry

The final version of this paper was published as:
Fischer, J.-P., & Thierry, X. (2022). Boy's math performance, compared to girls', jumps at age 6 (in the ELFE's data at least). *British Journal of Developmental Psychology, 40*(4), 504-519. https://doi.org/10.1111/bjdp.12423

Boy's math performance, compared to girls', jumps at age 6 (in the ELFE's data at least)

Abstract. The mathematics achievement discrepancy between girls and boys, with its subsequent occupational consequences, is an issue that has received considerable attention in the literature. It is often referred to as the "math-gap" and favors boys. A major component of the explanation of this gap resides in determining its age of onset. We analyze here data from more than ten thousand (cross-sectional study) and two thousand (longitudinal study) French students aged 4 to 7 years, tested in the framework of the Etude Longitudinale Française depuis l'Enfance (ELFE). The results allow to precisely determine the age of onset, since the gender difference, non-existent (or even slightly in favor of girls) in kindergarten (4-5 years), is clearly in favor of boys in first grade (6-7 years). They could therefore provide an important element in the controversial debate on the origin of gender-differentiated performance in mathematics.

Keywords: mathematics, gender, math-gap, kindergarten, first-grade, stereotype.

Data availability statement: The data that support the findings of this study are available on request from the ELFE CADE (https://pandora.vjf.inserm.fr/public). The data are not publicly available due to legal restrictions.

Acknowledgements: The ELFE study was carried out jointly by INED, INSERM, EFS, SPF, INSEE, DGS, DGPR, DREES, DEPS, and CNAF, with the support of MESRI and INJEP. The study received state aid coordinated by ANR in respect of the PIA with the reference of ANR-11-EQPX-0038.

Statement of Contribution

What is already known on this subject? Knowledge about the mathematics gap, favoring boys, is controversial. The gap is sometimes considered nonexistent or disappearing in society. Yet, the underrepresentation of girls in STEM remains an important economic problem today.

What does this study adds? This research provides the precise age at which the mathematics gap favoring boys develops. At this age, between kindergarten and first grade, something critical happens. Future studies should examine what might be promoting the development of such a gap, giving that there is no symmetrically evolving language gap (as girls always outperform boys).

1. Introduction

The gender difference in math performance, also known as the "gender gap in math" (Entwisle, Alexander, & Olson, 1994), is an important issue for both educational and economic policy. Because it predominantly favors one gender—boys (see Andreu et al., 2021, in France; Mullis et al., 2016, in countries with comparable educational system)—the math-gap can be seen as an expression of gender inequality in school and society (Anghel et al., 2020; Guiso et al., 2008; Gevrek, Gevrek, & Neumeier, 2020). Because it favors adult men, both young (OECD, 2020) or older (Jenkins et al., 2011), it is seen as responsible for women's underrepresentation in science, technology, engineering, and math future professions, although the girls' comparative advantage in reading can also explain this underrepresentation (Breda & Napp, 2019; Delaney & Devereux, 2021). The search for the origin and reason of the development of the math-gap during childhood or adolescence appears then, potentially, a necessary and major approach to expect avoiding it.

In the relatively recent past, the debate was around an innate versus societal origin of the gender difference in math performance. Today, the idea of a completely innate origin is difficult to defend. Firstly, because math is a fundamentally symbolic, abstract topic (Fischer, 2013) and therefore inaccessible to babies and toddlers. Secondly, because even non-symbolic math, when tested on toddlers, does not lead to an advantage for boys, as both Spelke (2005) and Kersey et al. (2018) pointed out. Consequently, the origin of the gap should not be before the age 3 or 4 years. As, on the other hand, the 2011 Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K, see Cimpian et al., 2016) revealed a clear and significant math-gap favoring boys in first grade, it seems appropriate to investigate its origin between the ages of 4 and 7.

At this age range we found research that investigated gender-differentiated mathematical activity in children, but only indirectly. For example, Thippana et al. (2020) followed 97 parent-child dyads for 6 months (M_{age} = 3 years 11 months, at the beginning of the study). They found that parents of boys were more likely to talk about numbers than parents of girls, but only during non-math activities. However, empirical research directly comparing boys' versus girls' mathematics performance, which is numerous in elementary or secondary school (e.g., Mullis et al., 2016; Pina et al., 2021; Rosselli et al., 2009), is rather scarce in children attending kindergarten.

Recent empirical research in China, in kindergarten and first grade, focused on only two numerical tasks, symbolic number comparison and estimation on a number line, and used a Bayesian factor approach to analyze the results (Zhang et al., 2020). In their Experiment 1, Zhang et al. found no substantial gender difference in the comparison task at either grade level; they did find strong evidence in favor of boys in the estimation task in kindergarten, but not in first grade. In their Experiment 2, they found no substantial evidence in favor of gender difference. These results, with two very specific tasks and a respectable but nonetheless limited number of participants (less than 100 in each of the three samples), do not allow us to draw any major conclusions as to the origin of the gender math-gap and even its existence.

In the United States there is the already mentioned ECLS-K study (Cimpian et al., 2016; Fryer & Levitt, 2010), replicated twice, in 1999 and 2011. Each of the ECLS-K studies initially involved approximately 20000 children. We are limiting ourselves here to a brief summary of the 2011 ECLS-K version (Cimpian et al.), not only because it is the most recent, but also because the table reported by Fryer and Levitt (p. 216) shows that boys were significantly older than girls, by almost 0.5 month. This age difference, at such a young age, may have biased the comparison between genders. The 2011 ECLS-K study included 2455 boys (M_{age} = 68.89 months) and 2601 girls (M_{age} = 68.29 months). The math assessments were adaptive and administered at four time points: Fall and Spring in kindergarten, Spring in first and second grade. Although boys performed slightly better than girls in both kindergarten sessions, the differences were not significant. The math-gap, in favor of boys, became clear and significant in first grade, and persisted and even increased in second grade.

Another research, by Kersey et al. (2018), investigated three domains of numerical development: numerosity perception, cultural trained counting and formal and informal mathematics. As suggested earlier, we do not consider infants' perception of numerosity to be a true test of mathematics. Similarly, the length of the count list and the "Give N" task are important but insufficient tasks for studying the math-gap. We therefore focused on the formal and informal mathematics investigated with the Test of Early Mathematics Ability (TEMA, 3rd edition) on 133 boys and 142 girls (M_{age} = 5.45, range: 3.07-7.92) by Kersey et al.. In this test, boys scored slightly but not significantly higher than girls, mean = 32.32 versus 30.04. Unfortunately, the more precise study of gender-differentiated informal and formal math scores as a function of age is biased by a data analysis problem resulting, for example, in 3-year-old boys having a higher proportion correct than some 7- and 8-year-old children. For a developmental psychologist or educator, this is difficult to understand.

Recent data from the French Longitudinal Study since Childhood (Enquête longitudinale française depuis l'enfance [Elfe]: Charles et al., 2020), make it possible to study the math-gap—its presence and possible emergence—in children aged 4 to 7 years. Precisely, we will investigate whether this gap emerges at the end of kindergarten. If this is the case, the math-gap in favor of boys should be present in first grade but not in kindergarten (intermediate section). This is our main (two-fold) hypothesis.

However, because comparison can help in interpreting data, we also report, in parallel, the result of the French language study in the same child samples. Gender differences in emerging language skills are well established (for review, see Rinaldi et al., 2021). For example, Eriksson et al. (2012) found such differences in 10 non-English language communities (including France) on 13783 one or two year olds. In a continuity of ages, the first ECLS-K cohort (see above), comparing 8182 girls with 8701 boys, found that girls have somewhat stronger literacy skills when they enter kindergarten, and also learn slightly more than boys during the kindergarten year (Ready et al., 2005). Thus, we can hypothesize that girls' performance in language will be more or less uniformly higher than that of boys. This is our secondary hypothesis, whose confirmation, if added to that of the main hypothesis, should constrain the explanatory theories of the latter.

2. Method

2.1. Participants

Over 18000 newborns were recruited across metropolitan France in 2011 as part of the Elfe study. The recruitment method is fully described in Charles et al. (2020). Five and 7 years after recruitment, most children were, respectively, attending the intermediate section of the kindergarten and first grade of the primary school in their place of residence. Consequently, in general there was just one Elfe child per class. The experimental attrition inherent in any long-term longitudinal study increased substantially through the strict anonymity of the children (whose teachers were unaware that they were part of the Elfe cohort before being informed) and teachers volunteering to administer and code the tests. This explains why there are just 2633 Elfe participants for whom we have complete data on all items submitted to the pupils, in kindergarten or first grade. The children's mean age was 57.28 (*SD* = 2.86) and 79.74 (*SD* = 2.80) months old, in kindergarten and first grade, respectively; 50.25% of them are girls.

To avoid isolating the Elfe child, initially the only one concerned by the test, the Elfe research designers asked the child's teacher, where possible, to simultaneously test three other classmates closest in age to the Elfe child. For these non-Elfe children, information is not always available for sex and age. Because the aim of our present research is the comparison of school-performance between genders, all non-Elfe children with unknown sex were excluded from the non-Elfe sample, both in kindergarten and first grade. Because the non-Elfe children in kindergarten are different from the non-Elfe children in first grade (with very few exceptions), we refer to the study of non-Elfe children as cross-sectional, which distinguishes it from the longitudinal study of Elfe children. The non-Elfe kindergarten (*N* = 9093) and first grade (*N* = 8184) samples had a mean age of 58.02 (*SD* = 3.15) and 80.40 (*SD* = 3.11) months, and included 51.51% and 51.81% girls, respectively.

2.2. Tasks and items

The cross-sectional and longitudinal studies' designs described above imply that the tasks and items were the same in both studies. However, they differ between kindergarten (intermediate section) and first grade. For each of the four tests (2 grades x 2 domains), content validity was addressed by determining the degree of correlation between an individual test item and the other test items (combined).

2.2.1. In kindergarten

The math subtest initially consisted of 26 items regrouped in 6 tasks (see Figure 1). However, the two last of the quotities comparison items were excluded due to low correlation (< .20) with all other items and close to chance responses (Fischer & Thierry, 2021a). Thus, the math subtest was reduced to 24 items. Content validity analysis in the combined samples (cross-sectional and longitudinal) yielded acceptable correlations of each item with all the others (combined), with the 24 *r*s ranging from 0.20 to 0.54.

As shown in Figure 1, the items concern both symbolic and non-symbolic numbers (for a differential analysis of the two types of numerical knowledge, see Fischer and Thierry, 2021a). Despite the small number of items, it is important to see that the inclusion of two types of items is consistent with current research describing children’s number and quantity development based on two systems. One, primitive—the approximate number system (ANS)—is shared with animals, whereas the other, more symbolic, requires knowledge of the number word sequence and can therefore only be transmitted through the child's cultural environment. These two systems, including their interactions, have been intensively studied over the past decade (e.g., Elliott et al., 2019; Ouyang et al., 2021).

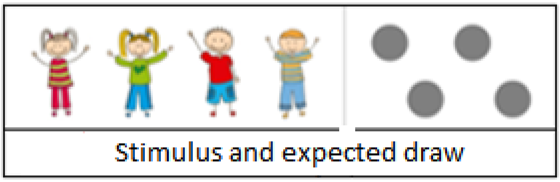
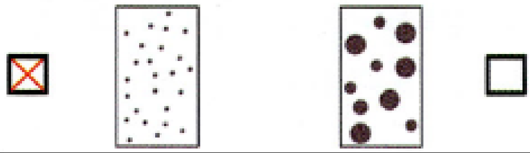
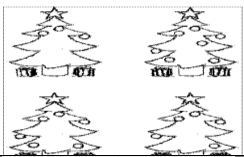
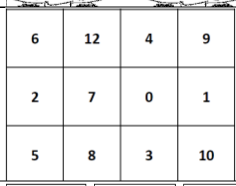
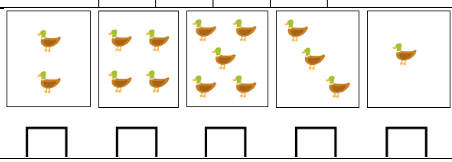
Task[a]	*n* items	implied number[a]	example of item
Match through drawing	3	4, 6, and 5	
Draw result	6	1+1+1, 2+2, 1+1+1+1, 3+2, 3-2, (1+1+1)-2[b]	The teacher drops counters into an opaque cup one by one or simultaneously
Compare quotities[c]	4	7 < 9, 25 > 10, 56 > 47, 59 > 25	
Recognize a stated number	3	three, five, seven	
Recognize written stated digits (or 10)	5	four, six, five, ten, seven	
Count and write the number	5	2, 4, 5, 3, 1	

Figure 1. Description of the tasks and items used in the kindergarten math-test.

[(a)] The task, items and numbers concerned are reported in order of appearance in the test. [(b)] The additive or subtractive notations indicate that the counters, or groups of counters, are added or removed one after the other. [(c)] Quotity is the property of being or representing a particular number of discrete things (as distinct from an amount or size).

The language subtest consisted of 35 items grouped into 4 tasks (see Figure 2). The items were adapted from a well-studied French kindergarten assessment battery (Labat et al., 2013). The 35 items relate to three major predictors of later reading success: letter name knowledge, phonological skills and vocabulary (Ecalle & Magnan, 2015). Content validity analysis yielded correlations slightly lower than in math, the 35 *r*s ranging from .18 to .46[1].

[1] This is a first reason why the item with a .184 correlation was left in the data, a second reason being that responses to this item considerably exceeded responses by chance, 66% vs. 25%.

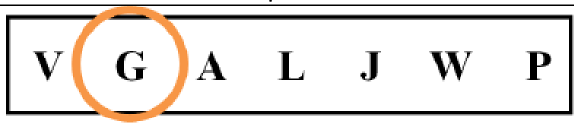
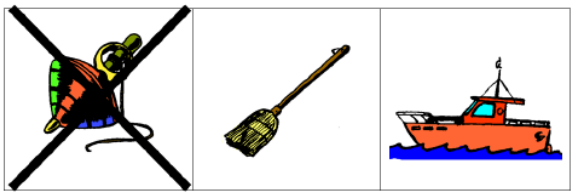
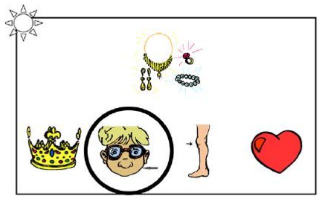
Task	*n* items	material	example of item
Circle the letter named	10	G, P, R, V, D, B, C, J, Q, T	 Recognize (by circling) the letter G
Cross out the word without the common unit[(a)]	9	[ba], [Yo], [Em], [ΣT], [ka], [bi], [be], [v], [zT]	 Recognize the sound [ba] by crossing out the intruder
Circle the new word after deleting the initial syllable[(b)]	6	joue, lait, riz, tronc, dos, pie	 Recognize the image of the cheek (joue, [Fu]) after removing [bi], in [biFu]
Circle the word you heard	10	fly, dive, sled, acorn, raft, deliver, chimpanzee, tulip, hanger, cabbage	 Recognize the word fly (by circling it)

Figure 2. Description of the tasks and items used in the kindergarten language-test.

[(a)] A non-French speaker should know that the sound [ba] is common to the French names of a brush (*balai)* and a boat (*bateau*), but not present in the name at top (*toupie*).[(b)] The picture above shows “bijoux”. By removing [bi] we get “joux” = “joue” = [Fu] phonetically = “cheek” in English.

2.2.2. In first grade

The math subtest consisted of 32 items divided in 5 tasks. In the first task, children had to write the results of 8 dictated mental computations: 3 + 7, 2 + ? = 5, 6 – 2, 9 – 8, 8 : 2, 2 * 9, 56 – 50, 62 – 10. In the second task, they had to solve 6 read story problems by writing down the numerical answer: for example, “I bought two packages of four candies. How many candies do I have altogether?”. In the third, fourth, and fifth tasks, they had to count 4 sums of money from visualized bills and coins, such as 20€ + 10€ + 5€ + 1€, to complete an empty space in 6 series of six numbers, such as 10 – 8 – ? – 4 – 2 – 0, and to put the sign <, =, or > in 8 comparisons of numbers, such as 70 to compare with 58, respectively.

Content validity was addressed in a sample of all children with usable data. All 32 correlations of an item with all others (combined) are between 0.34 and 0.64. These correlations, neither too small nor too large, are fully satisfactory.

The language subtest consisted of 45 items divided in 4 tasks. In the first task, children had to draw 10 collections of circles, each matching the number of sounds they had just heard; for example after hearing the pseudo-word “plar”, the expected answer was to draw 4 circles. In the second task, after the experimenter had read a story, they had to answer 10 comprehension questions by checking the image (among 3, with an extra box for not knowing) that answered the question concerned. In the third task, they had to read a story by themselves, and answer 10 comprehension questions, by checking the short written answer (among 3, with an extra box for not knowing) that answered the question concerned. In the fourth task, they were successively presented with 15 images (e.g., the image of a cochon) and had to choose the correct writing (out of 5) of the pictured object (e.g., here “cochon”, in competition with 4 distractors: cachon, cocher, cnocho, côchon).

Content validity analysis yielded the 45 correlations of one item with all others (combined) ranging from .15 to .51. The minimum correlation .15 appears somewhat low, but given its high statistical significance ($p < .001$) and the large number of correlations calculated (n = 45), the item concerned by this low correlation was nevertheless included in our analyses.

2.3. Procedure

At both school-levels tested, the teachers were provided with four booklets for pupils, a guide on conducting the tests, and a grid for coding the result in the last quarter of the school year. A pilot test was led a year earlier to collect the teachers' observations and to adjust the materials, instructions and assessments. The children's results were corrected by the teachers, who then sent them on to the Elfe coordination unit.

In general, the Elfe children were tested with three of their fellow pupils as suggested by the Elfe team. In addition to a booklet, the children were equipped with a pencil, an eraser, and, in kindergarten, five color crayons. The math subtest lasted about 25 minutes, and the language subtest a little longer (about 35 minutes), at a different time of the day or week. Both subtests were administered during school time.

Teacher participation was always voluntary. They were informed that the test was not a screening or diagnostic test. They asked the children not to answer aloud and not to copy from their peers.

2.4. Performance assessment

For the present analyses, a binary assessment, correct (= 1 point) or incorrect (= 0 point), was used for all items, except in kindergarten for the number writing items. For the latter, an otherwise correct mirror written response earned 0.5 point (these mirror writings have been analyzed in Fischer & Thierry, 2021b). Simply adding up these points leads to four global scores—Gmath4 (out of 24), Glang4 (out of 35), Gmath6 (out of 32), and Glang6 (out of 45)—for children ages 4-5 or 6-7, in math and language.

In addition, the a priori grouping of items into tasks allows the calculation of four task scores—Tmath4, Tlang4, Tmath6 and Tlang6—, by averaging the points obtained in each task. These task scores, initially between 0 and 1, as well as the global scores were transformed in scores out of 100 for better readability and possible interpretation in percent correct.

2.5. Statistical analyses

Half of our main hypothesis, that there is not gender difference in math in kindergarten, relies on not rejecting the null hypothesis (H_0). It seems therefore appropriate to compute the Bayes Factor (BF_{10}). The BF_{10} indeed measures the relative evidence of the model M_1 (boys and girls differ) over the model M_0 (no gender difference). To allow a comparison between the ages of boys and girls, which is expected not to differ, and a comparison with language, the BF10 was then systematically calculated. To interpret BF_{10}, we used the categories of evidence against H_0 from Kass and Raftery (1995), with a slightly modified verbal description: for BF_{10} from 1 to 3, anecdotal, from 3 to 20, substantial, from 20 to 150, strong, and $BF_{10} > 150$, very strong. For $BF_{10} < 1$, which is evidence in favor of H_0, the critical values must be inverted, that is, 0.333, 0.050, and 0.007. The R package BayesFactor (Morey & Rouder, 2021) was used to calculate BF_{10} with a prior rscale = $\sqrt[2]{2}/2$.

To take into account the variance introduced by the tasks, as well as the inter-individual differences among the participants, a Linear Mixed effects Model (LMM) was conducted to model the task scores of participants in first grade of the longitudinal study.

This grade and study were chosen because they permit the inclusion of participants' global kindergarten scores among the variables. These scores are known to be a significant predictor of children's first-grade reading and math outcomes, even the best predictor when kindergarten letter and number recognition was compared to other predictors, such as parental stress and children's behavior (Bramlett, Rowell, & Mandenberg, 2000). The inclusion of the global score in kindergarten allows then to test its predictor power of the score in first grade.

The LMM included participant and task as grouping factors. As fixed effects in the model, we considered the participants' gender (boys vs. girls), their standardized global score in kindergarten, their centered age (in months), as well as the interaction of the two first predictors. As random effects, we included random intercepts per participant and per task. The model was estimated with the lme4 function (Bates et al., 2015) of the R statistical program (R Core Team, 2021).

3. Results

3.1. Cross-sectional study

Table 1, fourth column, shows that there was no significant gender difference in age. With respect to our two-fold main hypothesis, an unexpected observation appears in the Math column: the math-gap is significantly (albeit weakly: $|d| = 0.131$) in favor of girls in kindergarten. The BF_{10} very strongly supports this gender difference but because it is in favor of girls, it can been see (at minima) as excluding a boy's advantage. This, added to the very strong support of a math-gap in favor of the boys in first grade, where the boy's advantage is significant ($|d| = 0.256$), clearly confirms our two-fold main hypothesis. In language, the advantage of the girls is significant and very strongly supported by the BF_{10}, albeit small (both $|d| < 0.12$), at the two school-levels.

Table 1. Cross-sectional comparison of the mean performance difference between boys and girls in math and language (French) at two grades or ages

Grade	Gender and difference	*N*	Mean age in months (*SD*)[(a)]	Mean score on 100 (*SD*)	
				Math	Language
Kindergarten (intermediate section)	Boys	4409	58.05 (3.16)	76.88 (18.65)	64.92 (18.95)
	Girls	4684	58.00 (3.15)	79.24 (17.33)	66.80 (17.84)
	p (t-test)	-	.446	< .001	< .001
	d (Cohen)	-	0.017	- 0.131	- 0.102
	BF_{10}	-	0.032	> 150	> 150
First grade	Boys	3944	80.43 (3.10)	74.66 (21.00)	66.43 (16.42)
	Girls	4240	80.36 (3.11)	69.30 (20.89)	68.34 (15.92)
	p (t-test)	-	.340	< .001	< .001
	d (Cohen)	-	0.022	0.256	- 0.118
	BF_{10}	-	0.027	> 150	> 150

[(a)] The values in this column were calculated with slightly lower *N*s because the age of some NElfe children was not specified (for 688 of them in kindergarten and 770 in grade 1).

Figure 3 visualizes the medians (bold horizontal lines), the two other quartile lines (horizontal sides of the boxes), and the minimum and maximum (the whiskers) when the outliers are excluded[2]. Focusing on the left side of the figure, it appears that in math, the quartile lines for boys clearly exceed the corresponding lines for girls in first grade, whereas the median, first quartile, and minimum lines for girls slightly exceed the corresponding lines for boys in kindergarten. The right-hand of Figure 3 shows that girls performed only slightly higher than boys at the two age- or school-levels in language.

[2] Outliers are arbitrarily defined (by default in R) as deviating by more than 1.5*IQR (Interquartile range) from the median.

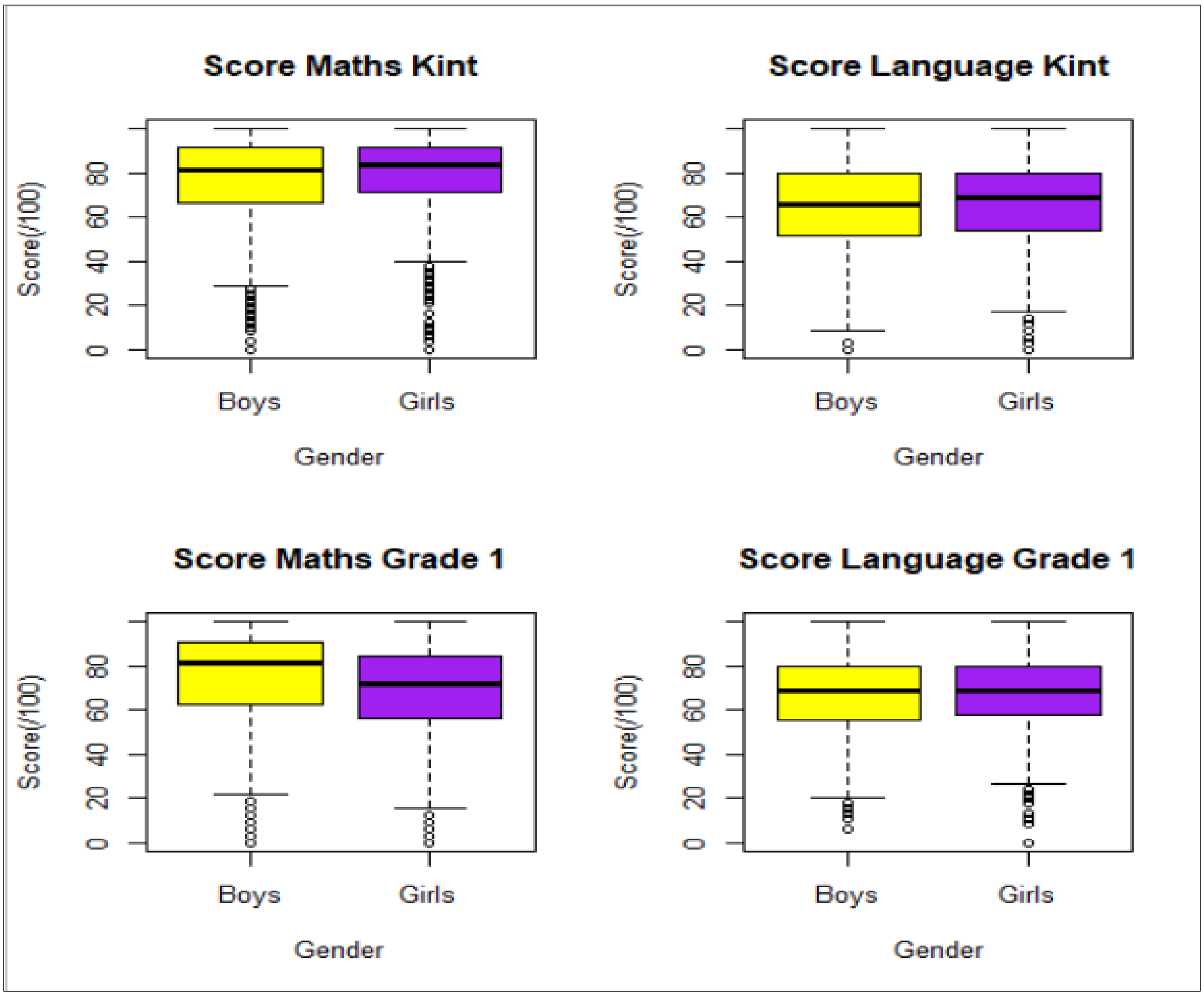


Figure 3. Boxplots of the gender difference in the cross-sectional study, in two domains (math and language) and two school-levels (intermediate kindergarten and first grade)

Figure 3 suggests a slight ceiling effect in the math performance in kindergarten. This ceiling could be responsible for the lower performance of the boys (as a group), as the higher-performing boys were not able to improve their performance. We examined such a possibility: there were 261 girls (5.6%) who performed at ceiling (Math4 score = 100), while only 220 boys (5.0%) performed at ceiling, a non-significant difference, $\chi^2(1) = 1.54$, $p = .21$. Thus, we can rule out this hypothesis.

3.2. Longitudinal study

3.2.1. Non-regression statistics

Table 2, for the participants in the longitudinal study, fundamentally corroborates the results of the statistics for the participants in the cross-sectional samples. Whereas the math-gap is not significant in kindergarten (H_0 is substantially supported by the BF_{10}), it is anew significant in favor of boys, small to median in size ($|d|$ = 0.306), and very strongly supported by the BF_{10} in first grade. In language, the girls' advantage is significant in first grade (H_1 is substantially supported by the BF_{10}); it is also significant in kindergarten but H_0 is anecdotally supported by the BF_{10}.

Table 2. Longitudinal comparison of the mean performance difference between boys and girls in math and language (French) at two grades or ages

Grade	Gender and difference	*N*	Mean age in months	Mean score on 100 (*SD*)	
				Math	Language
Kindergarten (intermediate section)	Boys	1310	57.36(2.88)	80.80 (16.87)	68.76 (17.75)
	Girls	1323	57.21 (2.84)	81.73 (15.22)	70.09 (17.04)
	p (t-test)	-	.199	.136	(<) .050
	d (Cohen)	-	0.051	-0.058	-0.076
	BF_{10}	-	0.108	0.131	0.413
First grade	Boys	1310	79.79 (2.80)	79.46 (18.23)	69.97 (15.02)
	Girls	1323	79.68 (2.80)	73.70 (19.32)	71.81 (14.91)
	p (t-test)	-	.297	< .001	.002
	d (Cohen)	-	0.041	0.306	- 0.122
	BF_{10}	-	0.073	> 150	3.194

Figure 4 illustrates our two-fold main hypothesis in math. There is no clear gender difference in kindergarten, whereas the minimum, first quartile, median, and third quartile lines for boys clearly exceed those corresponding for girls. In language, girls performed only slightly higher than boys at the two school-levels, as shown by the minimum, first quartile and median lines.

Figure 4, like Figure 3, suggests a slight ceiling effect in the math performance in kindergarten. There were 89 girls (i.e. 6.7%) and 96 boys (7.3%) who performed at ceiling, a non-significant difference, $\chi^2(1) = 0.36$, $p = .55$.

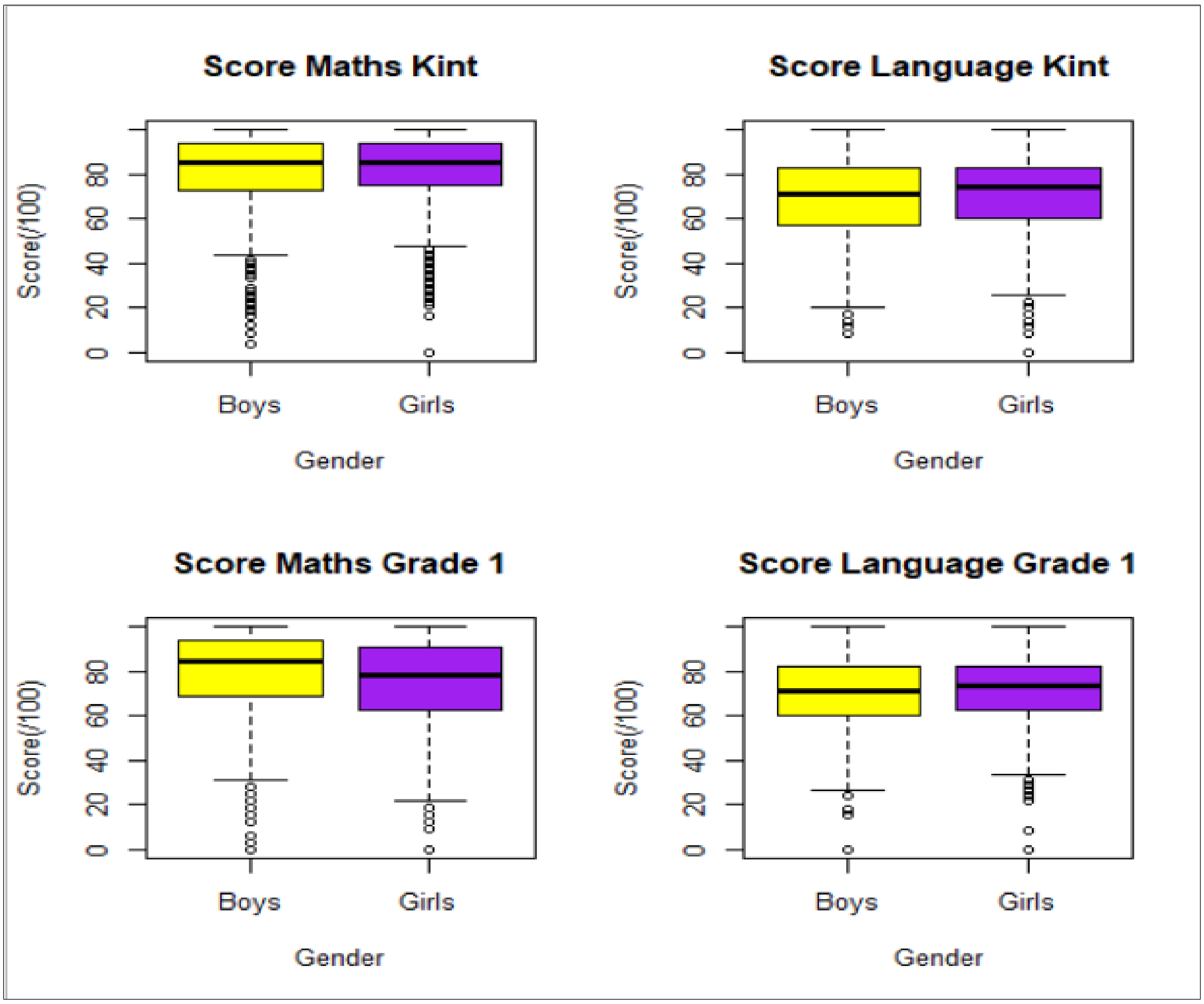


Figure 4. Boxplots of the gender difference in the longitudinal study, in two domains (math and language) and two school-levels (intermediate kindergarten and first grade)

3.2.2. Mixed effects regressions

As described in the Method-Statistical analyses section, two mixed effects regression of Tmath6 and Tlang6 scores were performed. Models with an increasing number of (nested) predictors, beginning with the null model were computed. For Tmath6, beginning with the null model Mmath6.0 and then successively adding the predictors Sex, Gmath4, Age, and Sex*Gmath4 (interaction), we obtained the models Mmath6.1, Mmath6.2, Mmath6.3, and Mmath6.4, respectively, each of which significantly improving the previous model (all four chi-square tests were significant at the .01 level). For comparability, the same predictors were used for the Tlang6 score. The resulting Mmath6.4 and Mlang6.4 models are summarized in Table 3, and the models and the results of their step-by-step changes are detailed in the Supporting information (appended).

Table 3 shows that the coefficients of Sex predictor are significant, favoring boys in math and girls in language. These coefficients, and their significance, are consistent with the lesser impact of the Sex predictor in language than in math in the non-regression statistics analyses. Table 3 also shows the significant impact of the global score at age 4, both in math and language, and its interaction with Sex, suggesting that the effect of Sex is not constant over all levels of the G-score.[3]

[3] This interaction expresses, for example, that the average difference in math between boys and girls is higher in the bottom quartile than in the top quartile of Gmath4 scores (but always in favor of boys).

Table 3. Mixed effects linear regression fit by REML of the task-scores

	In math			In language		
Fixed effects	Estimate	*SE*	*p*[(a)]	Estimate	*SE*	*p*[(a)]
Intercept	79.52	4.64	< .001	71.78	5.82	.001
Sex (ref: Boys)	-6.44	0.66	< .001	1.30	0.52	.013
G-score: age 4	8.33	0.45	< .001	5.26	0.37	< .001
Age	0.33	0.12	.007	0.03	0.09	> .743
Sex * G-score	1.73	0.67	.009	1.04	0.52	.046
Random effects		*SD*	*n*		*SD*	*n*
Participant		14.53	2633		9.00	2633
Task		10.32	5		11.61	4
Residual		19.61			19.79	

[(a)] Results of t-tests using Satterthwaite's method

4. Discussion

4.1. The main hypothesis

The observed pattern of a clear math-gap in favor of boys in first grade, whereas such a gap is inexistent in kindergarten (intermediate section), with even girls performing higher than boys in the cross-sectional study, has an important consequence. If we accept that the math-gap develops with some regularity, and not in a zig-zag fashion, the age of its origin should be indicated by the jump around 5-6 years.

This main result was obtained in both the cross-sectional and longitudinal study. There is, in fact, a double convergence in our research results. First, the non-regression statistics, whether frequentist or Bayesian, from both cross-sectional and longitudinal studies are consistent. This is particularly noteworthy since these two studies involved completely different participants. Second, these statistics and linear mixed effects modeling in the longitudinal study are also consistent. The result of the LMM shows that taking into account some important variables, such as the level of performance in kindergarten (intermediate section), does not weaken, or at least not vanish the measure of gender's influence in first grade.

The main result is also enhanced by the comparison with language. In this domain, our secondary hypothesis of a general advantage of girls was clearly confirmed in the cross-sectional study, as well as in the mixed effects analysis, and somewhat less clearly (but in no way contradictorily) in the longitudinal study. This comparison with language thus reinforces the interest of our main result of the emergence of a gender gap in math around age 5 by suggesting its specificity.

Research on older students suggests that gender gaps in mathematics favoring males appear mainly, if not solely, at the top of the distribution of achievement scores (Casey & Ganley, 2021; Hedges & Nowell, 1995; Stoet & Geary, 2013). In addition, Robinson and Lubienski (2011) made a similar observation in kindergarten. The combination of the slight ceiling effect, which we observed in the kindergarten math data, and this specificity of the math-gap may have prevented us from finding a gap in favor of boys in kindergarten. However, a gender-differentiated ceiling effect was not confirmed in our analysis about Figures 3 and 4. Thus, the possible non-constancy of the gender gap in the distribution of achievement scores does not seem to have affected our result.

4.2. Comparison with other results

4.2.1. Convergent results

First, we can note that the pattern of results from the 2011 ECLS-K study, described in the introduction, seems consistent with our present observation. Indeed, Cimpian et al. (2016) observed a nonsignificant math-gap pattern in favor of boys in kindergarten with children slightly older than those in our intermediate kindergarten section, and a confirmation of the gap in first grade. This pattern of results, however, would require a risky extrapolation to suggest that the trend observed at 5.5 years, and its increase at 6 or 7 years, did not yet exist at 4.5 years.

A possible causal statement of our main finding appears in the article by Cvencek, Meltzoff, and Greenwald (2011, p. 766) when they stated that “the math–gender stereotype is acquired early and influences emerging math self-concepts prior to ages at which there are actual differences in math achievement”. Given that these authors studied U.S. students in grades 1-5, the influence of gender stereotype in math on differences in math achievement, as evidenced in first grade, could however be bidirectional when these differences are already present in first grade (as we here demonstrated).

Contini, Di Tommaso and Mendolia (2017) showed that girls systematically underperform boys in selected grades 2-10 in Italy. In addition, they observed that the gender gap increases with children’s age. Although these data and observation are consistent with ours, they left open the question of the age of appearance of the gap.

4.2.2. Diverging results

Because of our clear demonstration that boys outperform girls in math in first grade, our results are at odds with all those that place the emergence of the gap later than first grade. For example, Fryer and Levitt (2010) concluded that gender gaps in math arise as early as third grade, suggesting that this grade is the age of onset of the gap.

Our result in first grade is also at odds with the general conclusion by Gunderson et al. (2012, p. 163), that “there is no longer a gap between boys and girls on math achievement tests “. In fact, this claim is not only discordant with our result but also with the researches that have studied the gap and generally found it (e.g., Cimpian et al., 2016; Contini et al., 2017; Perez-Mejias et al., 2021). However, it must be recognized that publication bias may have favored the publication of this latter research.

Kersey et al.’s (2018, p. 6) assertion, by reference to two articles by Hyde and colleagues (Hyde, Fennema & Lamon., 1990; Hyde & Linn, 2006), “that school test performance differences in mathematics between boys and girls are nonexistent or trivial during elementary school” is also rather at odds with our results. Of course, the differences we found are not very large. But d = 0.306 (see Table 2), which falls between "small" and "medium" according to Cohen's (1988) nomenclature, does not seem a negligible difference.

Johnson et al. (2022) found that boys outperformed girls in kindergarten quite significantly, $t(376) = 2.735$, $p = .007$, $d = 0.28$. But, the children were, on average, 6 years old. They are therefore better compared with our first graders than with our intermediate kindergarten children. The discrepancy with our results is therefore much smaller than it seems at first glance.

4.2.3. A non-empirical study

The research by Perez-Mejias et al. (2021) seems a priori promising because the authors state in their abstract that the purpose of the study “was to track the gender scoring gap in mathematics from kindergarten to grade 12”. Thus, it is very astonished to find only repeated measures of achievement in grades 4, 8, 10 and 12. Presenting then, in their Figure 2, the observed data (in grades 4-12) and the predicted data (in kindergarten), together and without distinction, is troubling: the reader is led to believe that in kindergarten, boys performed 0.15 *SD* (or 0.18 in the conditional model) higher than girls, when in fact no performance was measured in kindergarten. The authors concluded that their “findings provide evidence gender differences first occur before children enter the school system”. However, we prefer real data to model predictions, no matter how sophisticated the models are.

4.3. Limitations

The Elfe longitudinal sample is probably not representative from the French population of this age. The children cannot be considered as representative because their parents had to participate in the test on a voluntary basis. Fortunately, the additional recruitment of 9093 and 8184 participants in kindergarten and first grade, respectively, for the cross-sectional study, partially remedies this representativeness problem. This is because these participants were recruited solely on the basis of their age.

A common remark about gender differences is that they are small (Beller & Gafni, 1996; Casey & Ganley, 2021; Else-Quest, Hyde, & Linn, 2010; Hyde & Linn, 2006). We have already argued that a *d* of about 0.30 is not negligible. But the majority of our *d*’s are rather around 0.10 (see Tables 1 and 2). This is the case for our somewhat original finding that girls significantly outperform boys in math, in the cross-sectional study, in kindergarten ($p < .001$, $d = -0.131$). However, the BF_{10} very strongly supports this difference and, importantly, its direction is not in favor of boys.

A possible reservation about our analyses comes from the choice not to exclude some apparently extreme participants (see the "outlying" points in Figures 3 and 4). This non-exclusion is based mainly on our statistical conviction that it is the mass of data that should remedy certain accidents (Benzécri, 1968), and secondarily on the practical difficulty of finding and applying an operational and non-arbitrary exclusion criterion. In addition, Figures 3 and 4 show that outliers affect—and do not clearly differentiate— boys and girls.

5. Conclusion

We investigated the early development of the gender gap in mathematics in two large cross-sectional samples and a longitudinal sample of children aged 4-5 years and 6-7 years from the French Elfe study. If the data and analyses are reliable, our double result in the comparison of school mathematics performance between gender—a non-existing math-gap in favor of boys at age 4- to-5-years (or even an advantage for girls) and a clear advantage for boys among 6- to 7-year-olds—may be unique in the large literature on differential performance in mathematics by gender. Indeed, as Kersey et al. (2018) note, if there are intrinsic gender differences in aptitude, then differences in quantitative and mathematical abilities should emerge early in human development. Our analyses, like the much more complete analyses by Kersey et al. suggest that they do not appear early in human development. But most importantly, the present research goes one step further. It shows how rapidly, in the space of one or two years, the mathematical gender gap favoring boys develops. This additional step—knowing that something critical happens in the last year of kindergarten (ages 5 to 6)—should stimulate research into what might promote the development of such a gap.

References

Andreu, S., Ballereau, M. A., Ben Ali, L., Bret, A., Dos Sanstos, R., et al. (2021). Evaluation de début de sixième 2021 : premiers résultats. Paris : MENESR & DEPP.

Anghel, B., Rodríguez-Planas, N., & Sanz-de-Galdeano, A. (2020). Is the math gender gap associated with gender equality? Only in low-income countries. *Economics of Education Review, 79*(C). doi: 10.1016/j.econedurev.2020.102064

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01.

Beller, M., & Gafni, N. (1996). The 1991 international assessment of educational progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology, 88,* 365-377. doi: 10.1037/0022-0663.88.2.365

Benzécri, J. P. (1968). La place de l'a priori. In *Encyclopaedia Universalis, 17*, 11-24.

Bramlett, R., Rowell, K., & Mandenberg, K. (2000). Predicting first grade achievement from kindergarten screening measures: a comparison of child and family predictors. *Psychology in the Schools, 7*(1), 1-9.

Breda, T., & Napp, C. (2019). Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *PNAS, 116*(31), 15435-15440. doi: 10.1073/pnas.1905779116

Casey, B. M., & Ganley, C. M. (2021). An examination of gender differences in spatial skills and math attitudes in relation to mathematics success: A bio-psycho-social model. *Developmental Review, 60,* 100963. doi: 10.1016/j.dr.2021.100963

Charles, M. A., Thierry, X., Lanoë, J. L., Bois, C., Dufourg, M. N., Popa, R., Cheminat, M., Zaros, C., & Geay, B. (2020). Cohort Profile: The French national cohort of children ELFE: birth to 5 years. *International Journal of Epidemiology, 49*(2), 368-369j. doi: 10.1093/ije/dyz227

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*(4). doi: 10.1177/2332858416673617

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Taylor & Francis (second edition).

Contini, D., Di Tommaso, M. L., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review, 58,* 32-42. doi: 10.1016/j.econedurev.2017.03.001

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development, 82*(3), 766–779. doi: 10.1111/j.1467-8624.2010.01529.x

Delaney, J. M., & Devereux, P. J. (2021). High school rank in Math and English and the gender gap in STEM. *Labour Economics, 69,* 101969. doi: 10.1016/j.labeco.2021.101969

Ecalle, J., & Magnan, A. (2015). *L'apprentissage de la lecture et ses difficultés.* Paris : Dunod (second edition revised).

Elliott, L., Feigenson, L., Halberda, J., & Libertus, M. E. (2019). Bidirectional, longitudinal associations between math ability and approximate number system precision in childhood. *Journal of Cognition and Development, 20*(1), 56-74. doi: 10.1080/15248372.2018.1551218

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*(1), 103–127. doi: 10.1037/a0018053

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1994). The gender gap in math: its possible origins in neighborhood effects. *American Sociological Review, 59*(6), 822-838. doi: 10.2307/2096370

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez-Pereira, M., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M., & Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology, 30*(2), 326-343. doi: 10.1111/j.2044-835X.2011.02042.x

Fischer, J. P. (2013). L'abstraction réfléchissante : une spécificité des mathématiques ? *Rééducation Orthophonique, 255,* 7-18.

Fischer, J. P., & Thierry, X. (2021a). Are differences between social classes reduced by non-symbolic numerical tasks? Evidence from the ELFE cohort. *British Journal of Educational Psychology, 91*(1), 286-299. doi: 10.1111/BJEP.12363

Fischer, J. P., & Thierry, X. (2021b). Miswriting (especially mirror writing) of the digits: An ecological assessment using ELFE data. *Journal of Cognitive Education and Psychology, 20*(1), 3-17. doi: 10.1891/JCEP-D-19-00053

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*(2), 210-240. doi: 10.1257/app.2.2.210

Gevrek, Z. E., Gevrek, D., & Neumeier, C. (2020). Explaining the gender gaps in mathematics achievement and attitudes: The role of societal gender equality. *Economics of Education Review, 76,* 101978. doi: 10.1016/j.econedurev.2020.101978

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164-1165. doi: 10.1126/science.1154094

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. *Sex Roles, 66*(3), 153-166. doi: 10.1007/s11199-011-9996-2

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41-45. doi: 10.1126/science.7604277

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin, 107*(2), 139-155. doi: 10.1037/0033-2909.107.2.139

Hyde, J. S., & Linn, M. C. (2006). Gender similarities in mathematics and science. *Science, 314*(5799), 599-600. doi: 10.1126/science.1132154

Jenkins, A., Ackerman, R., Frumkin, L., Salter, E., & Vorhaus, J. (2011). *Literacy, numeracy and disadvantage among older adults in England.* University of London: Institute of Education.

Johnson, T., Burgoyne, A. P., Mix, K. S., Young, C. J., & Levine, S. C. (2022). Spatial and mathematics skills: Similarities and differences related to age, SES, and gender. *Cognition, 218,* 104918. doi: 10.1016/j.cognition.2021.104918

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795. doi: 10.2307/2291091

Kersey, A. J., Braham, E. J., Csumitta, K. D., Libertus, M. E., & Cantlon, J. F. (2018). No intrinsic gender differences in children's earliest numerical abilities. *npj Science of Learning, 3,* 12. doi: 10.1038/s41539-018-0028-7

Labat, H., Farhat, S. L., Andreu, S., Rocher, T., Cros, L., Magnan, A., & Ecalle, J. (2013). Évaluation des connaissances précoces prédictives de l'apprentissage de la lecture en grande section de maternelle. *Revue Française de Pédagogie, 184,* 41-54. doi: 10.4000/rfp.4219

Morey, R. D., & Rouder, J. N. (2021). Package 'BayesFactor'. https://CRAN.R-project.org/package=BayesFactor.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). TIMSS 2015 international results in mathematics. Retrieved, September 18, 2021, at: http://timssandpirls.bc.edu/timss2015/international-results/

OECD (2020). Do gender gaps in reading and mathematics evolve between childhood and adulthood? *Adult Skills in Focus, n° 13.* doi: 10.1787/6db40c8d-en.

Ouyang, X., Yang, Y., Zhang, X., & Zhang, Q. (2021). Longitudinal relations between the approximate number system and symbolic number skills in preschool children. *Journal of Experimental Child Psychology, 212*, 105254. doi: 10.1016/j.jecp.2021.105254

Perez-Mejias, P., McAllister, D. E., Diaz, K. G., & Ravest, J. (2021). A longitudinal study of the gender gap in mathematics achievement: evidence from Chile. *Educational Studies in Mathematics, 107,* 583-605. doi: 10.1007/s10649-021-10052-1

Pina, V., Martella, D., Chacón-Moscoso, S., Saracostti, M., & Fenollar-Cortés, J. (2021). Gender-based performance in mathematical facts and calculations in two elementary school samples from Chile and Spain: An exploratory study. *Frontiers in Psychology, 12*, 703580. doi: 10.3389/fpsyg.2021.703580

R Core Team (2021). *R: A language and environment for statistical computing.* Vienna (Austria): R Foundation for Statistical Computing.

Ready, D. D., LoGerfo, L. F., Burkam, D. T., & Lee, V. E. (2005). Explaining girls' advantage in kindergarten literacy learning: Do classroom behaviors make a difference? *The Elementary School Journal, 106*(1), 21-38. doi: 10.1086/496905

Rinaldi, P., Pasqualetti, P., Volterra, V., & Caselli, M. C. (2021, on line). Gender differences in early stages of language development. Some evidence and possible explanations. *Journal of Neuroscience Research*. doi: 10.1002/jnr.24914

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct gender gap development cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268-302. doi: 10.3102/0002831210372249

Rosselli, M., Ardila, A., Matute, E., & Inozemtseva, O. (2009). Gender differences and cognitive correlates of mathematical skills in school-aged children. *Child Neuropsychology, 15*(3), 216-231. doi: 10.1080/09297040802195205

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist, 60*(9), 950-958.

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

Thippana, J., Elliott, L., Gehman, S., Libertus, K., & Libertus, M. E. (2020). Parents' use of number talk with young children: Comparing methods, family factors, activity contexts, and relations to math skills. *Early Childhood Research Quarterly, 53*(4), 249-259. doi: 10.1016/j.ecresq.2020.05.002

Zhang, T., Chen, C., Chen, C., & Wei, W. (2020). Gender differences in the development of semantic and spatial processing of numbers. *British Journal of Developmental Psychology, 38*(3), 391-414. doi: 10.1111/bjdp.12329

Supporting information for the mixed effects regressions

[Section 3.2.2 of the article: “Boy’s math performance, compared to girls’, jumps at age 6 (in the ELFE’s data at least)”]

The models were fitted, using the *lmer* function of the *lme4* package, with REML, but were refitted with ML for their step-by-step changes. The non-evident names of the variables are explained in the article. The order of the predictors in the language analysis was taken from the math analysis.

Table S1: Description of the Models fitted in Maths

Name	Description, with the R syntax (package lme4)
Mmath6.0	Tmath6 ~ (1 | Participant) + (1 | Task)
Mmath6.1	Tmath6 ~ Sex + (1 | Participant) + (1 | Task)
Mmath6.2	Tmath6 ~ Sex + Gmath4 + (1 | Participant) + (1 | Task)
Mmath6.3	Tmath6 ~ Sex + Gmath4 + Age + (1 | Participant) + (1 | Task)
Mmath6.4	Tmath6 ~ Sex + Gmath4 + Age + Sex * Gmath4 + (1 | Participant) + (1 | Task)

Table S2: Step-by-step changes of the Models in Maths

Name	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Mmath6.0	4	119997	120027	-59994	119989			
Mmath6.1	5	119938	119975	-59964	119928	61.0374	1	5.6e-15 ***
Mmath6.2	6	119249	119294	-59619	119237	690.4522	1	< 2.2e-16 ***
Mmath6.3	7	119244	119297	-59615	119230	7.0402	1	0.007970 **
Mmath6.4	8	119240	119300	-59612	119224	6.8014	1	0.009108 **

Significance codes: 0 ‘***’; 0.001 ‘**’; 0.01 ‘*’; 0.05 ‘.’

Table S3: Description of the Models fitted in Language

Name	Description, with the R syntax (package lme4)
Mlang6.0	Tlang6 ~ (1 | Participant) + (1 | Task)
Mlang6.1	Tlang6 ~ Sex + (1 | Participant) + (1 | Task)
Mlang6.2	Tlang6 ~ Sex + Glang4 + (1 | Participant) + (1 | Task)
Mlang6.3	Tlang6 ~ Sex + Glang4 + Age + (1 | Participant) + (1 | Task)
Mlang6.4	Tlang6 ~ Sex + Glang4 + Age + Sex * Glang4 + (1 | Participant) + (1 | Task)

Table S4: Step-by-step changes of the Models in Language

Name	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Mlang6.0	4	94843	94872	-47417	94835			
Mlang6.1	5	94835	94872	-47413	94825	9.3149	1	0.002273 **
Mlang6.2	6	94388	94431	-47188	94376	449.7788	1	< 2.2e-16 ***
Mlang6.3	7	94390	94440	-47188	94376	0.1023	1	0.749121
Mlang6.4	8	94388	94446	-47186	94372	3.9917	1	0.045725 *

Significance codes: 0 ‘***’; 0.001 ‘**’; 0.01 ‘*’; 0.05 ‘.’

References

R Core Team (2021). *R: A language and environment for statistical computing.* Vienna (Austria): R Foundation for Statistical Computing.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01.