Metacognition and Learning (2021) 16:157–178
https://doi.org/10.1007/s11409-020-09243-7

Gender differences in confidence during number-line estimation

Michelle L. Rivers[1] · Charles J. Fitzsimmons[1] · Susan R. Fisk[2] · John Dunlosky[1] · Clarissa A. Thompson[1]

Received: 8 June 2020 / Accepted: 1 September 2020 / Published online: 7 October 2020

Abstract

Prior research has found gender differences in spatial tasks in which men perform better, and are more confident, than women. Do gender differences also occur in people's confidence as they perform number-line estimation, a common spatial-numeric task predictive of math achievement? To investigate this question, we analyzed outcomes from six studies ($N = 758$ girls/women and boys/men with over 20,000 observations; grades 1–5 and adults) that involved a similar method: Participants estimated where a provided number (e.g., ¾, 37) was located on a bounded number line (e.g., 0–1; 0–100), then judged their confidence in that estimate. Boys/men were more precise ($g = .52$) and more confident ($g = .30$) in their estimates than were girls/women. Linear mixed model analyses of the trial-level data revealed that girls'/women's estimates had about 31% more error than did boys'/men's estimates, and even when controlling for precision, girls/women were about 7% less confident in their estimates than were boys/men. These outcomes should encourage researchers to consider gender differences for studies on math cognition and provide pathways for future research to address potential mechanisms underlying the present gender gaps.

Keywords Confidence judgments · Estimation · Gender differences · Number lines

Imagine two first graders who just rated their confidence on a number-line estimation trial in which they estimated where 25 goes on a 0 to 100 number line. Although both first graders greatly overestimated the location of the number on the line, one child was more confident in the estimate than the other. Such differences in confidence can influence the children's engagement with math, because higher confidence is related to greater self-efficacy and persistence when tasks are difficult. Now imagine that the more confident child is a boy and the less confident child is a girl, and that such differences are not just isolated to these children.

✉ Michelle L. Rivers
mlrivers3@gmail.com

[1] Department of Psychological Sciences, Kent State University, Kent, OH, USA

[2] Department of Sociology, Kent State University, Kent, OH, USA

Does a gender gap occur in which girls are less confident than boys when they are engaged in math tasks such as number-line estimation? Given that such a gap could have negative short- and long-term consequences for educational and career outcomes, discovering whether such a gender gap exists is critical so that future research can develop interventions to minimize it.

The present research evaluates the extent to which gender differences arise in confidence on number-line estimation, a task which taps the fundamental ability to estimate numerical magnitude (and is predictive of future math achievement; e.g., Bailey et al. 2014; Booth and Siegler 2006, 2008; Fazio et al. 2014; Fuchs et al. 2010; Geary 2011; Schneider et al. 2018; Siegler 2016; Siegler et al. 2011, 2012; Siegler and Thompson 2014; Tosto et al. 2018). In this task, participants are (1) asked to estimate where a provided number (e.g., 25) falls on a horizontal number line and (2) asked to judge their confidence in their estimate – an example is presented in Fig. 1. In the remainder of the Introduction, we explain why exploring gender differences in confidence for number-line estimation is important and why such differences in confidence may occur. Finally, we describe our analytic approach for estimating gender differences in this domain.

Why Investigate Gender Differences in Confidence for the Number-Line Estimation Task?

Our reasons for focusing on gender differences in confidence for number-line estimation arise from the possibility that gender differences will occur in number-line estimation performance. Thus, to motivate our interest in confidence, we begin by first considering why (and whether) gender differences occur in number-line estimation performance. According to Halpern's (2004) *cognitive-process taxonomy*, considering the component processes that underlie task performance is essential for understanding the degree to which gender differences may occur in performance. Whereas women tend to perform better on some tasks that require verbal skills, men tend to perform better on some tasks that require spatial skills. Consistent with this taxonomy, Halpern et al. (2007) concluded that a performance gap favoring boys emerges in tasks that require spatial ability, such as the mental-rotation task (for reviews, see Lauer et al. 2019; Levine et al. 2016; Voyer et al. 1995; Voyer et al. 2017).

To the extent that space and number are intertwined in the number-line estimation task – the focal task in our analyses – one might anticipate gender differences due to the inherent spatial characteristics of the task. That is, to perform successfully on number-line estimation, people must map their mental representation of a numerical magnitude onto a one-dimensional spatial representation (i.e., the number line). In fact, some researchers have hypothesized that people possess a *mental* number line in which smaller numbers are oriented on the left side of space and larger numbers are oriented on the right side of space. For example, when preschoolers were shown a "hiding box" in which compartments in the box were verbally labeled with increasing numbers from left-to-right, they were quicker to find the object in a "finding box" that was similarly labeled with increasing numbers from left-to-right as compared to when they tried to find hidden objects in a finding box that was verbally labeled from right-to-left (Opfer et al. 2010). Further, in the *spatial numerical association of response codes effect* (SNARC effect; Dehaene et al. 1993; Hubbard et al., 2005), people (who read left-to-right) tend to respond faster when smaller numbers require a left-hand button press and bigger numbers require a right-hand button press.[1] Bull et al. (2013) found this SNARC effect to be

[1] Although, see Colling et al. (2020) for a failure to replicate a priming paradigm in which participants were quicker to detect targets on the right side when they were preceded by large numbers.

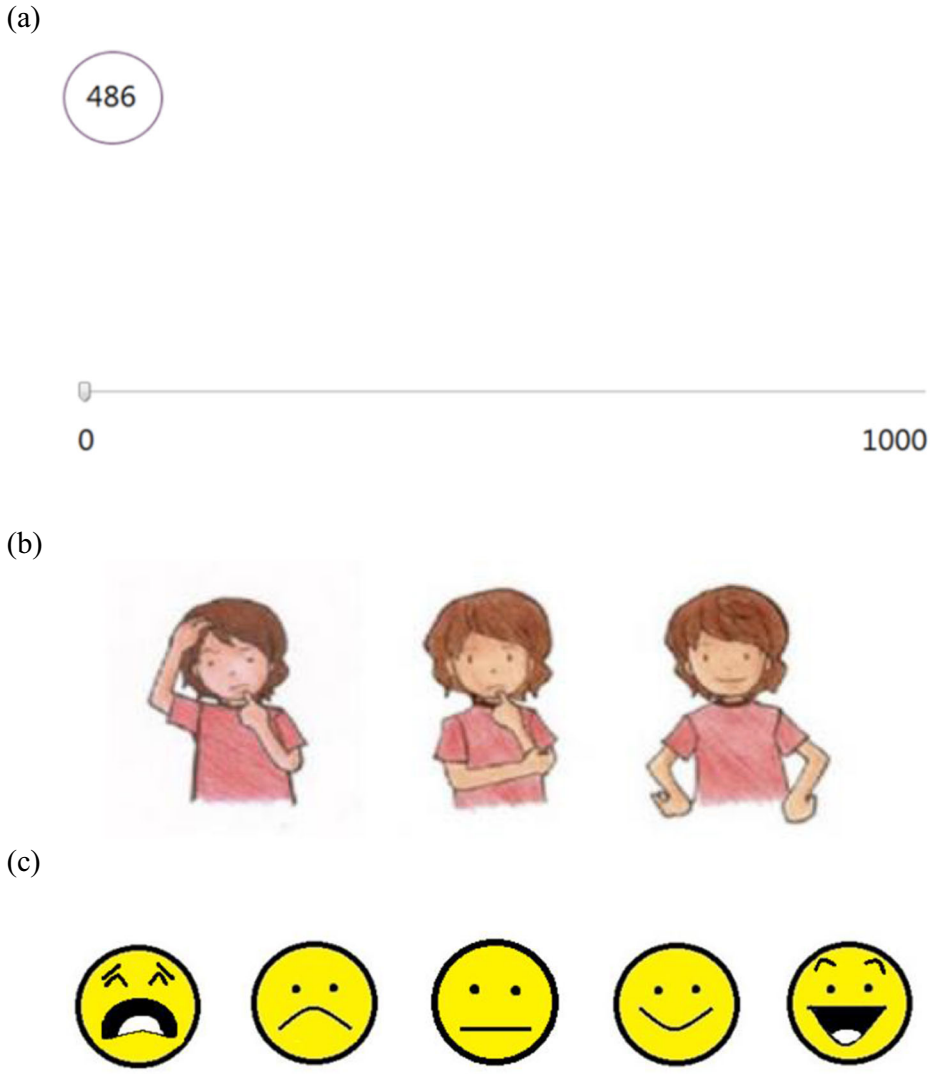


Fig. 1 Sample Number-Line Estimation Trial with Confidence Judgment. *Note*. Children used the slider (**a**) to estimate the target number on the number line, then used the pictures at the bottom to make their confidence judgments: (**b**) 3-point confidence judgment scale (from Fitzsimmons et al. n.d.; adapted from Hembacher and Ghetti 2014); (**c**) a 5-point confidence judgment scale (from Wall et al., n.d.)

stronger in men than in women, suggesting that the association between numerical and spatial representation may be more pronounced in men.

Consistent with the above rationale, gender differences have been observed in number-line estimation performance across development and for various numerical scales, with medium effect sizes on average (Bull et al. 2013; Gunderson et al. 2012; Hutchinson et al. 2019; LeFevre et al. 2010; Reinert et al. 2017; Thompson and Opfer 2008). For example, in a recent study of gender differences in tasks tapping basic numerical skills (e.g., counting ability, number comparison, arithmetic, etc.), Hutchinson et al. (2019) found that number-line estimation was the only task for which boys performed better than did girls. Specifically, the

authors investigated gender differences in number-line estimation performance in the 0–100 and 0–1000 range across approximately 1400 children in the Netherlands in grades 1–6 and reported an overall medium effect favoring boys (Cohen's $d = 0.39$ and 0.59 for the 0–100 and 0–1000 number-line scales, respectively). To foreshadow, such gender differences in number-line estimation performance also occurred in the present research, which motivated our focus on confidence.

Will Gender Differences Occur in Confidence About Number-Line Estimation Performance?

Given such gender differences in number-line estimation performance, will gender differences also occur in confidence? Answering this question is important for a couple reasons. First, if girls/women are less confident in their estimation performance (as compared to boys/men), differences in confidence could partly be contributing to the differences in performance (for an example in the context of the mental rotation task, see Estes and Felker 2012). In particular, low confidence could lead to a lack of task persistence. Second, given that gender differences do occur in number-line estimation performance, any gender differences in confidence could arise because people's confidence tracks their performance. For instance, if confidence judgments are perfectly accurate (i.e., aligned with performance), then the confidence judgments would naturally be lower for those who perform worse on the task. Note, however, that prior research investigating the accuracy of confidence judgments for the number-line estimation task has demonstrated that their accuracy is far from perfect (although accuracy is not entirely on the floor; e.g., Wall et al. 2016). Thus, any gender differences in confidence could arise from (somewhat) accurate monitoring of performance or from gender biases in making confidence judgments. Thus, if gender differences occur in confidence, will they remain when gender differences in performance is statistically controlled?

To answer our focal questions, we used a method that was illustrated in our opening scenario. When performing this task, people's confidence can be measured by having them judge how well they performed on each trial; these *retrospective confidence judgments* (henceforth, confidence judgments for brevity) refer to people's confidence that their estimate accurately represents the location of that number's magnitude on the line. As with other metacognitive judgments (e.g., judgments of learning), confidence judgments are not based on direct access to how precisely numbers are represented in memory. Rather, theories of metacognition distinguish between two types of information that can be used as a basis for judgments (e.g., Koriat 1997; Koriat and Ackerman 2010; Koriat and Levy-Sadot 1999). *Theory-based* judgments are informed by people's naive beliefs about learning or their perceptions about their own abilities. In contrast, *experience-based* judgments are informed by on-line monitoring during task performance. In the number-line estimation task, both of these factors could influence confidence judgments about estimation performance. For example, people may base their judgments on their beliefs about their ability to perform in a math/spatial domain (i.e., self-efficacy; Bandura 1977), their attitudes about whole numbers and fractions (e.g., Sidney et al. 2019), or perceived difficulty of the number-line estimation task (i.e., theory-based factors). Or, people may base their judgments on inferences about cues available when they are responding on a specific trial (i.e., experience-based factors), such as having the experience of familiarity with to-be-estimated numbers (e.g., Fitzsimmons et al. 2019; Fitzsimmons et al. 2020). Next, we consider how each of these factors may contribute to gender gaps in confidence.

One theory-based factor – self-efficacy – is most pertinent to the present research as it may be a basis for confidence judgments. In particular, children who believe they tend to perform well on math or spatial tasks may be more likely to make higher confidence judgments following number-line estimation compared to those who feel less confident in their abilities. On the one hand, when asked to rate their overall math or spatial ability, boys/men tend to make higher ratings than do girls/women (e.g., Ariel et al. 2018; Ganley and Lubienski 2016; Syzmanowicz and Furnham 2011). In addition, attitudes about math tend to be more negative for girls/women compared to boys/men, and women also report being more anxious about math than do men (for a review, see Hyde et al. 1990). On the other hand, young girls earn better grades in mathematics throughout elementary school (e.g., Dwyer and Johnson 1997; Halpern et al. 2007; Kenney-Benson et al. 2006; Kimball 1989; Marshman et al. 2018) and hence this early recognition of their good grades in math may lead them to be more confident in their performance as compared to boys. Thus, given the potential role of self-efficacy in confidence judgments, the available evidence may lead one to expect that a gender difference (in either direction) will emerge in confidence judgments for number-line estimation.

In addition to theory-based factors such as self-efficacy, confidence judgments can also be constructed from experience-based inferences, such as how quickly a person responds to a given question (with quicker responses producing more confidence, e.g., Benjamin et al. 1998) or a person's overall familiarity with the numbers they are estimating (with more familiar numbers producing higher confidence, e.g., Fitzsimmons et al. 2020). To the degree that gender differences occur in such experiences (e.g., in familiarity), those could result in gender differences in confidence judgments as well.

To date, few investigations are available about gender differences in trial-by-trial confidence judgments for math tasks, and that evidence is mixed. Even when controlling for performance, gender differences have been observed in trial-by-trial confidence judgments for various math and spatial tasks (e.g., mental rotation, paper folding, solving arithmetic problems), with boys being more confident than girls (e.g., Ariel et al. 2018; Boekaerts and Rozendaal 2010; Cooke-Simpson and Voyer 2007; Estes and Felker 2012). However, other studies have reported no gender differences in confidence for math tasks, such as solving arithmetic problems or number-discrimination decisions (e.g., Baer and Odic 2019; Nelson and Fyfe 2019). Thus, whether gender differences will occur in confidence for this task remains an open question. As important, we also investigated whether gender differences occur in confidence when task performance (i.e., estimation precision) is controlled. This analysis is critical because when gender differences occur in performance and in confidence, then the former differences in performance could (appropriately) be producing the gender differences in confidence.

Method Overview

Replicating prior research, our preliminary analyses (described below) revealed gender differences in number-line estimation performance favoring boys ($g = .52$). Accordingly, the present investigation will provide answers to two critical questions that have not been previously addressed:

1. Do gender differences exist in confidence for number-line estimation precision? And, if the answer to this question is "yes", then:
2. Do gender differences in confidence remain when controlling for gender differences in estimation precision?

In answering these questions, we realize that making accurate claims about gender differences is challenging, as research on gender differences is often plagued by publication bias (e.g., Nelson 2014). To avoid this issue, we conducted analyses using all available data (that we are aware of) on the number-to-position number-line estimation task in which participants also made trial-by-trial confidence judgments. To obtain a comprehensive list of empirical studies that met this inclusion criteria, we undertook a systematic search of the literature through the APA PsychINFO database and Google Scholar. The database search involved using the keyword "number line" in combination with the terms "metacognition" and "confidence judgments." No date restrictions were applied during the literature search, which concluded on August 10, 2020. The database searches yielded 9 unique hits. After screening these papers against our criteria, only Wall et al. (2016) and Fitzsimmons et al. (2020) survived. We also conducted a forward search (i.e., citations of original articles) for these two articles and screened all citations against our criteria; this did not yield any additional articles that met our criteria. We also included additional datasets from our laboratories, some of which, at the time of writing, have not yet been published or were undergoing peer review.

These studies were particularly well-suited for avoiding publication bias as none of them were designed to examine gender differences in number-line estimation performance and none (prior to this analysis) had involved analyzing gender differences. Our analyses include data from six studies (10 separate experiments) that examined participants' ability to estimate target numbers on a number line and their confidence in those estimates (Fig. 1; Feltner and Thompson n.d.; Fitzsimmons et al. 2019; Fitzsimmons et al. 2020; Fitzsimmons et al. n.d.; Wall et al., n.d.; Wall et al. 2016). In sum, the analyses included 758 participants (339 boys/men and 419 girls/women) ranging in age from first grade to adults. All of these studies received approval from Kent State University's Institutional Review Board, and informed consent was obtained for all participants. For full transparency, we note that the Principal Investigator (PI) was the same for all of these studies.[2] However, the PI worked with several different graduate and undergraduate research assistants, postdocs, and faculty collaborators to collect the data. Table 1 contains a summary of the descriptive information for the studies included in the analyses.

All experiments used a similar method in which participants made between 12 and 48 estimates for whole numbers or fractions on number lines. Six different numerical ranges were used: 0–10, 0–100, 0–1000, 0–100,000, and 1000–1 billion (for whole numbers), and 0–1 (for fractions). In some cases, participants made estimates in multiple numerical ranges. Each to-be-estimated number was presented one at a time on an individual sheet of paper or on a computer screen. Estimation precision was measured on a trial-level basis as *proportion absolute error* (PAE). In this paper, PAE is calculated with the following equation:

$$\mathrm{PAE} = |\text{Participant's estimate} - \text{Correct Answer}| / \text{Scale of number line} \quad (1)$$

For example, if a participant was asked to estimate the location of "70" on a 0–100 number line but marked the location corresponding to "90," the PAE would be computed as [(|90–70|)/

[2] Given data were all collected by the same lab, readers may be concerned that the same participants completed multiple experiments. Because we recruited from some of the same school districts each year, there is a (small) possibility that this occurred (e.g., 4th graders in 2016 and 5th graders in 2017). However, most of the data collected was separated by multiple years, and we can be sure that the same children did not participate in multiple studies - for example, we are certain that 1st and 2nd grade children from the Wall et al. (2016) dataset were not the same 1st and 2nd graders in the Fitzsimmons et al. (n.d.) dataset.

Table 1 Summary of Study-Level and Sample-Level Characteristics

Study	Sample size (*N*)	Grade	Percent Boys/Men	Number-Line Scale/s	Number of Estimations	Confidence Judgment Scale
Wall et al. (2016), Expt 1	18	1	50%	0–10; 0–100	18 per scale (36 total)	3 point
	24	2	50%	0–100; 0–1000	18 per scale (36 total)	3 point
	17	4	47.1%	0–1000; 0–100,000	18 per scale (36 total)	3 point
Wall et al. (2016), Expt 2	17	1	47.1%	0–10; 0–100	18 per scale (36 total)	3 point
	21	2	52.4%	0–100; 0–1000	18 per scale (36 total)	3 point
	16	4	31.3%	0–1000; 0–100,000	18 per scale (36 total)	3 point
Wall et al. (2016), Expt 3	18	1	44.4%	0–100	18	3 point
	31	2	41.9%	0–1000	18	3 point
	36	4	44.4%	0–100,000	18	3 point
Wall et al. (n.d.), Expt 1	20	2	55.0%	0–100; 0–1000	18 per scale (36 total)	5 point
Fitzsimmons et al.(n.d.), Expt 1	72	1	48.6%	0–1000	22	3 point
	38	2	39.5%	0–1000	22	3 point
Fitzsimmons et al. (n.d.), Expt 2	117	Adults	12.8%	1000–1 billion	22	100 point
Fitzsimmons et al. (2020), Expt 1	90	Adults	62.2%	0–1	44	4 point
Fitzsimmons et al. (2020), Expt 2	101	Adults	58.4%	0–1	12	4 point
Feltner and Thompson (n.d.), Expt 1	51	5	49%	0–1	20	3 point
Fitzsimmons et al. (2019)	58	4	44.8%	0–1000; 0–100,000; 0–1	18 per scale for whole numbers, 12 per scale for fractions (48 total)	5 point
	13	5	53.8%	0–1000; 0–100,000; 0–1	18 per scale for whole numbers, 12 per scale for fractions (48 total)	5 point

Note. Total $N = 758$; $k = 18$. One additional participant from Fitzsimmons et al. (n.d.) and another participant from Fitzsimmons et al. (2020) were excluded from analyses for choosing not to report their gender. The Principal Investigator was the same for all of these studies

100] = .20 (Eq. 1). Higher deviation indicates a less precise estimate. PAE can range between 0 and 1.

Immediately after estimating each number, participants rated their confidence for each estimate (Fig. 1). Confidence judgments were made on three-point (*not so sure*, *kind of sure*, or *really sure;* Wall et al. 2016.; Fitzsimmons et al. n.d., four-point (*not so sure, kind of sure,*

pretty sure, or totally sure; Fitzsimmons et al. 2020), five-point (Fig. 1; Feltner and Thompson, n.d.; Fitzsimmons et al. 2019; Wall et al., n.d.), and 100-point (Fitzsimmons et al. n.d.) scales. Adults self-reported their gender on demographic forms, and parents reported the gender of their children on parental permission forms.[3] For experiments that included multiple experimental phases, we only included data collected prior to when a manipulation was introduced. For example, Fitzsimmons et al. (n.d.) used a pretest-posttest design to investigate the effectiveness of interventions aimed at improving estimation performance; only pretest data were included in the current analyses.

We use two separate methods to conduct our analyses: (1) commonly used meta-analytic procedures in which we calculate standardized measures of effect sizes (Hedges' g), and (2) linear mixed models that use trial-level data. We first present the Hedges' g meta-analyses, in which we examine whether gender differences exist in estimation precision (performance) and confidence judgments (confidence). We then present the linear mixed model analyses. In these analyses, we (1) replicate the results from the Hedges' g meta-analyses and (2) examine whether gender differences in confidence exist controlling for estimation precision at the trial level.

Results

Hedges' g Analyses

Analytic Approach To calculate gender differences in performance and confidence, we conducted meta-analyses using standard procedures in which standardized effect sizes were calculated and then weighted by the number of participants.

We computed Hedges' g (Hedges 1981) as a standardized measure of effect size, which represented the standardized mean difference between boys/men and girls/women. For each sample, g was computed using the following equation:

$$g = (M_{\mathrm{m}} - M_{\mathrm{f}})/s_{\mathrm{pooled}} \tag{2}$$

In Eq. 2, M_{m} = the mean for boys/men, M_{f} = the mean for girls/women, and s_{pooled} = the pooled estimate of the population standard deviation. In these calculations, a g greater than 0 indicated an advantage for boys/men in precision or confidence, whereas a g less than 0 indicated an advantage for girls/women. For studies that had multiple experiments, each experiment contributed a unique effect size given that the settings, instructions, and other factors varied by experiment. If multiple effect sizes were available from the same sample (e.g., the same participants made estimates on multiple number-line scales, as in Wall et al. 2016), we pooled the means and standard deviations, and these pooled values were used to compute the composite effect sizes that were included in the analyses (for justification, see Card 2012; p. 192–193). In addition, for experiments that investigated number-line estimation across multiple grades, each grade contributed a separate effect size (given prior research showing that the precision of children's estimates improves with age; Siegler and Booth 2004; Siegler and

[3] For full transparency, we note that participants were asked to report their "sex" on these demographic forms. Because we are not making claims that any differences observed between men and women can be attributed solely to biological differences (e.g., differences in physical attributes between males and females; American Psychological Association 2012), we use the term "gender" throughout the paper and refer to boys/men and girls/women rather than males and females.

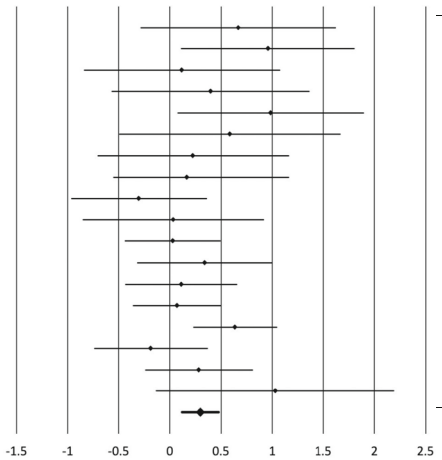
Opfer 2003; Siegler et al. 2009; Thompson and Opfer 2010). Ultimately, the six studies included in the meta-analyses produced 18 effect sizes composed of 758 participants (Table 1).

We used random-effects models for our meta-analyses (e.g., Hedges 1983; Hedges and Vevea 1998; Raudenbush 1994). Unlike fixed-effects models, which conceptualize only a single population mean effect size, random-effects models estimate a mean population effect size as well as the variability in effect sizes due to the population variability in effect sizes and hence are more appropriate to use when the effect sizes of the studies included in the analysis differ from each other (Borenstein et al. 2010). Random-effects models also allow for inferences that generalize beyond the samples included in the meta-analyses to a broader population of potential representative studies (i.e., *unconditional inferences*; Hedges and Vevea 1998). Because we anticipated variation in the effect sizes, random effects models were considered most appropriate for the overall analyses. For more details on how we conducted these analyses, refer to Appendix 1.

Gender Differences in Performance Our analyses revealed medium gender differences in number-line estimation performance favoring boys/men ($g = .52$; Appendix 2).

Gender Differences in Confidence A gender difference occurred in confidence favoring boys/men (Fig. 2). The overall weighted effect size was $g = .30$, 95% CI [.12, .47], $p = .002$. No significant heterogeneity was observed among the effect sizes, $Q(17) = 19.03$, $p = .33$. However, with 18 samples, we were only powered to detect large amounts of heterogeneity (Card 2012, p. 191). We also calculated I^2, which describes the percentage of total variation across studies that is due to heterogeneity rather than sampling variability alone (Higgins and Thompson 2002). According to Higgins et al. (2003), I^2 values of 25%, 50%, and 75% are associated with small, medium, and large amounts of heterogeneity, respectively. This index revealed a small, non-significant amount of heterogeneity among the effect sizes; $I^2 = 10.69\%$.

For comparison, we also conducted a fixed-effects model (for similar justification, see Lawson et al. 2018). The fixed-effects model resulted in a similar overall mean effect size, $g = .26$, 95% CI [.10, .41], $p = .002$.



Study, Expt, and Grade	Effect size	95% CI lower	95% CI upper
Wall et al. (2016), Expt 1, Grade 1	0.67	-0.28	1.62
Wall et al. (2016), Expt 1, Grade 2	0.96	0.11	1.80
Wall et al. (2016), Expt 1, Grade 4	0.12	-0.84	1.07
Wall et al. (2016), Expt 2, Grade 1	0.40	-0.56	1.36
Wall et al. (2016), Expt 2, Grade 2	0.99	0.08	1.89
Wall et al. (2016), Expt 2, Grade 4	0.59	-0.49	1.66
Wall et al. (2016), Expt 3, Grade 1	0.23	-0.71	1.16
Wall et al. (2016), Expt 3, Grade 2	0.17	-0.55	1.16
Wall et al. (2016), Expt 3, Grade 4	-0.30	-0.96	0.36
Wall et al. (n.d.), Expt 1, Grade 2	0.03	-0.85	0.91
Fitzsimmons et al. (n.d.), Expt 1, Grade 1	0.03	-0.43	0.49
Fitzsimmons et al. (n.d.), Expt 1, Grade 2	0.34	-0.31	1.00
Fitzsimmons et al. (n.d.), Expt 2, Adults	0.11	-0.43	0.65
Fitzsimmons et al. (2020), Expt 1, Adults	0.07	-0.36	0.50
Fitzsimmons et al. (2020), Expt 2, Adults	0.64	0.23	1.04
Feltner & Thompson (n.d.), Expt 1, Grade 5	-0.19	-0.74	0.36
Fitzsimmons et al. (2019), Expt 1, Grade 4	0.28	-0.23	0.80
Fitzsimmons et al. (2019), Expt 1, Grade 5	1.03	-0.13	2.19
Grand Mean	**0.30**	**0.12**	**0.47**

Fig. 2 Forest Plot of Effect Sizes for the 18 Samples Included in the Hedges' g Meta-Analysis of Confidence

Linear Mixed Model Analyses

Analytic Approach In addition to the Hedges' *g* meta-analyses, we also conducted linear mixed effects models (i.e., a specific type of multilevel model) so that data at the trial level (i.e., each estimate and its associated confidence judgment for a given participant) could be included in the analyses. This method increased our statistical power by allowing us to include 20,801 observations (nested within 758 participants) in the analyses and to control for performance on a given number-line estimate, which allowed us to answer the following question: Do gender differences exist in confidence, controlling for estimation precision? That is, are gender differences in confidence still observed when accounting for the fact that girls/women have lower estimation precision than boys/men?

We first replicated the effect of gender on estimation precision and confidence, then examined whether gender differences in confidence remained when estimation precision was statistically controlled. In these linear mixed effects models, trial-level observations (level-1 unit) are nested within participants (level-2 unit), which are then nested within experiments (level-3 unit).

Measure of Trial-Level Confidence Given that different scales were used to measure confidence in different experiments (i.e., some experiments used a 3-point scale, whereas others used 4-, 5-, or 100-point scales; Table 1), we converted each trial-level confidence judgment to a proportion (i.e., a 0–1 scale) using the formula: [(trial-level confidence judgment - 1) / (maximum confidence judgment value for the scale – 1)]. Thus, on a 3-point scale, a confidence judgment of 1 becomes 0 ([1–1]/[3–1] = 0), 2 becomes .50 ([2–1]/[3–1] = .50), and 3 becomes 1.0 ([3–1] / [3–1] = 1). Overall, the average trial-level confidence judgment was .653 (Table 2).

Gender Differences in Performance To summarize the main outcomes, we again found that girls/women were less precise in their number-line estimates than were boys/men (replicating the findings from the Hedges' *g* meta-analysis). The full analysis is presented in Appendix 3.

Trial-Level Gender Differences in Confidence, Controlling for Performance Using linear mixed-effects models to predict confidence, we again found that girls/women were slightly less confident than boys/men (replicating the findings from the Hedges' *g* meta-analysis). We also found that these gender differences remained when estimation precision was accounted for.

Table 3 presents two linear mixed effects models that predict trial-level confidence judgment proportion (on a 0–1 scale). Model A predicts confidence using only the participant's gender and Model B predicts confidence using both the participant's gender and the participant's trial-level PAE as a control.

Table 2 Trial- Level Summary Statistics for Linear Mixed Models

	Mean	*SD*	Min	Max
PAE	.145	.166	.000	.998
Confidence judgment	.653	.299	.000	1.00

Note. 20,801 trials nested within 758 participants. PAE = proportion absolute error. In total, 53.9% of participants were girls/women.

Table 3 Linear Mixed Models Predicting Trial-Level Confidence Judgments

	Model A Coeff. (*SE*)	Model B Coeff. (*SE*)
Fixed Effect Estimates		
Intercept	.688*** (.020)	.719*** (.020)
Participant's PAE on the trial	–	−.247*** (.012)
Woman	−.048** (.014)	−.038** (.014)
Random Effect Estimates		
U3 (Experiment)	.054 (.014)	.055 (.015)
U2 (Participant)	.182 (.005)	.180 (.005)
U1 (Trial)	.235 (.001)	.233 (.001)
AIC	881.0864	477.4147
BIC	920.8002	525.0713

Note. Trials nested within participant, which is nested within experiment. Model A predicts confidence using only the participant's gender and Model B predicts confidence using both the participant's gender and the participant's trial-level PAE as a control

**$p < .01$. ***$p < .001$

Model A replicates the findings from the Hedges' g meta-analyses on the effect of gender on confidence judgments. Using the fixed-effects estimates, girl's/women's confidence was estimated to be .048 points lower than boy's/men's ($p = .001$; Model A). The intercept of .688 ($p < .001$) represents the fixed-effects estimate of the average boy's/man's confidence judgment. Thus, the fixed-effects estimate of the average girls'/women's confidence judgment is .640 (.688–.048; see Model A). Thus, it appears that the average girl/woman was about 7% less confident than the average boy/man (.048 / .688 = .0697; Model A).

Given that boys/men were both more precise in estimating magnitudes (e.g., had lower PAE) and were also more confident in their estimates compared to girls/women, one possibility is that girls'/women's lower confidence reflects their poorer performance. That is, perhaps girls/women are just as confident as boys/men once their higher PAE is taken into account. Model B tests this hypothesis by adding a control for estimation precision (i.e., participant's trial-level PAE). This control reduced the magnitude of the gender coefficient from −.048 ($p = .001$, Model A) to −.038 ($p = .007$, Model B). However, gender remained a statistically significant predictor of confidence: Even when controlling for trial-level estimation precision, girls/women were .038 points less confident in their estimates than were boys/men.

Discussion

Recall the boy and girl estimating numbers in our opening vignette. Though neither was very precise in the location where they placed the number on the number line, confidence in estimation performance was higher for the boy than the girl. The vignette partly reflected the gender gaps revealed in the present research. In particular, we analyzed all available data on the number-line estimation task in which participants *also* made trial-by-trial confidence judgments. In meta-analyses of six studies with 758 participants (339 boys/men and 419 girls/women) and subsequent linear mixed models of trial-by-trial number-line estimates and confidence judgments, boys/men were more precise than girls/women in their number-line estimates. In addition, boys/men were more confident than girls/women, *even when* controlling for estimation precision. For the remainder of the Discussion, we review potential mechanisms, future directions, and implications of these findings.

Gender Differences in Number-Line Estimation Performance and Confidence

As compared to girls/women, why did boys/men make more precise number-line estimates? Although considerable debate exists over the causes of such gender differences when they are observed (e.g., Hyde 2014), we imagine that psychological (e.g., differences in math attitudes; Sidney et al. 2019), social (e.g., differences in early spatial experiences, such as exposure to spatial language, media, and toys; Caldera et al. 1989; Doyle et al. 2012; Pruden and Levine 2017; or gendered stereotypes about math and spatial ability, McGlone and Aronson 2006; Moè and Pazzaglia 2006), and possibly even biological factors (e.g., sexual dimorphism in the parietal cortex; Goldstein et al. 2001) could contribute to the gender differences observed in the number-line estimation task (as is the case for performance; e.g., Tosto et al. 2018). We leave it to future research to evaluate the contribution of these factors to the present gender differences observed.

As compared to girls/women, why were boys/men more confident in their trial-by-trial confidence judgments? As mentioned in the Introduction, confidence judgments can be influenced by multiple theory- and experience-based factors (e.g., Undorf et al. 2018). The observed gender gap in confidence could be explained by differences in judgment cue use by girls/women and boys/men. We consider two possibilities here. First, the gender gap may reflect a concomitant gap in people's self-efficacy for performing math tasks. In this case, as compared to boys/men, girls/women may in general believe they are less capable of solving math problems, which in turn could lower girl's/women's confidence judgments even when they performed the number-line estimation task with equal precision. Another contributing factor is based on recent evidence demonstrating that confidence judgments are more strongly related to people's familiarity with estimated numbers than they are to actual performance (Fitzsimmons et al. 2019; Fitzsimmons et al. 2020). The idea here is that people's familiarity with specific numbers influences their number-line estimation confidence, with greater familiarity producing greater confidence.

Although the present data do not provide a test of these hypotheses, both are testable and arguably could contribute to the gender gap in confidence on this task. That is, as compared to boys/men, girls/women may generally have less task-specific efficacy for number-line estimation (either because of its math component, spatial component, or both) and may also have less perceived familiarity with the numbers. In fact, lower self-efficacy may push some girls/women away from engaging in math tasks, which in turn could reduce their actual familiarity with those numbers. Most important, with minor adaptations of the methods used here, the degree to which these (and other) factors jointly contribute to this gender gap in confidence could be estimated. For instance, number-line estimation with trial-by-trial confidence judgments could be supplemented with self-report methodology (e.g., asking participants to self-explain why they rated their confidence as they did) or by collecting people's math self-efficacy and perceived familiarity with the numbers used in the task.

Implications for Education

Although we have characterized our effects of gender as small/medium according to statistical convention, the practical implications of these findings are currently unknown and hence important to explore (Rosenthal et al. 2000). For example, could a small gender difference on number-line estimation impact performance on high-stakes tests, such as the SAT and GRE? Performance on the number-line estimation task is critically important to help people know

whether they are "in the ballpark" when solving a math problem and may partially explain why number-line estimation is related to concurrent math ability (e.g., Fazio et al. 2014) and predictive of future math ability (e.g., Bailey et al. 2014). And, self-confidence about math ability has been found to be an important contributor to women's and men's course enrollment, major, and career choices (Correll 2001; Ellis et al. 2016). As such, gender differences for math and spatial tasks may be one of a myriad of reasons that women are less likely to pursue and persist in STEM fields (for other reasons, see Ceci and Williams 2011; Ceci et al. 2009).

Given that a gender gap exists for number-line estimation and such a gap has possible implications for persistence in math, what can be done to narrow it? Some interventions have been found to effectively improve estimation precision (Fitzsimmons et al. n.d.; Opfer and Siegler 2007; Opfer and Thompson 2008; Thompson and Opfer 2010). For example, one of the studies included in our analyses (Fitzsimmons et al. n.d.) investigated the effectiveness of three interventions for improving both performance and confidence. Participants first completed a pretest in which they made a set of number-line estimates and made confidence judgments after each estimate. They then completed an intervention, which involved either (1) studying correct worked examples, (2) studying incorrect worked examples, or (3) receiving corrective feedback on their pretest performance. On a posttest, performance was significantly improved (compared to pretest) for participants who received corrective feedback or studied correct worked examples. However, the magnitude of confidence judgments did not change from pretest to posttest for most groups. Thus, employing educational interventions aimed at improving people's *confidence* for number-line estimation remains a fruitful avenue for future research.

One fascinating possibility is that interventions aimed at increasing girls'/women's confidence in their task performance may benefit performance – and not only on the trained math task (i.e., the number-line estimation task in the present case) but also on other math tasks. The possibility itself derives from self-efficacy theory (Bandura 1977): People (in this case girls/women) who are less confident in their ability to perform a task will be less persistent to solve it accurately. For instance, if a student believes she will not perform well, she may not consider effective solutions or may respond without considering other possible solutions. In this case, poor self-efficacy is a self-fulfilling prophecy in that the belief in not doing well contributes to poorer performance. This possibility may also partly explain (as discussed next) why gender differences were smaller in confidence than in performance, as if the lower confidence by girls/women added to what would otherwise be a smaller deficit in number-line estimation performance. If this lower-confidence-undermines-performance hypothesis is correct, then interventions aimed at enhancing confidence should also improve performance.

Limitations and Future Directions

In the current study, we found gender differences in confidence, even when controlling for performance. We also found that the magnitude of the gender differences observed for confidence were smaller than those for performance ($g = .30$ versus .52). This may lead some readers to wonder what this suggests regarding gender differences in judgment accuracy. For example, are boys/men *over*confident in their number-line estimates compared to girls/women? Unfortunately, our data do not allow us to answer this question. To compute measures of absolute accuracy (e.g., degree of overconfidence), judgments must be made on the same scale as performance. In many of the experiments included in our analyses, CJs were made on a 3-, 4-, or 5-point scale, whereas PAE is calculated as the absolute deviation between the correct

location on the number line and the participant's estimate expressed as a proportion (i.e., divide by the range of the scale). Although we could have transformed the scales to be comparable (e.g., divided the 3-point CJ scale by 3 to yield a proportion to be consistent with PAE), the resulting index cannot be "unambiguously interpreted" (Dunlosky et al. 2016, p. 30). For example, we cannot be sure that a CJ rating of "2" on a 3-point scale corresponds to .66 (or 66% confidence) in a participant's mind.

Interpretation is made even more complicated by the fact that performance on the number-line task is calculated as PAE, a continuous measure of performance. To assess whether a person is over- or under-confident, we would need to establish that an estimate is correct or incorrect and compare this to a participant's confidence for their estimate. However, people are rarely if ever exactly correct (i.e., PAE of 0) on this spatial-numeric task. For instance, even adults who are experts in the 0–1000 numerical range exhibit some small PAE for the task (e.g., Thompson and Opfer 2010, Experiment 1). To make it easier to calculate over- and under-confidence, PAE could be dichotomized as correct vs. incorrect (e.g., within 10% of the correct location). However, there is no justified theoretical reason to dichotomize this continuous variable, and in doing so, researchers would lose important variability in individual differences in this measure. Because of these issues, assessing the degree of over- or under-confidence for the number-line estimation task will require future advances in measurement of judgments and performance (so they can be made on comparable scales) for this task.

Although we cannot calculate measures of absolute metacognitive accuracy, we were able to assess whether gender differences exist for *relative* accuracy, or the degree to which participants can discriminate between number-line estimates that are more (vs. less) precise. We calculated relative accuracy by computing a gamma correlation for each participants' CJs and PAE (reverse coded for ease of interpretation) across trials, then averaged these values across participants (Nelson 1984). We chose to report this measure of relative accuracy to be consistent with and comparable to our previous publications (Wall et al. 2016; Fitzsimmons et al. 2020; Fitzsimmons et al. n.d.). Gamma is a nonparametric correlation that does not assume equal intervals between levels of a measure, with values ranging from −1 to 1 (with values closer to 1 reflecting better discrimination, and values closer to 0 indicating very little ability to discriminate between the precision of estimates). Across all participants, we found the mean values of gamma were positive for both genders (boys/men: $n = 309$, $M = .18$, $SD = .31$; girls/women: $n = 400$, $M = .20$, $SD = .28$), which suggests that participants have some ability to monitor the accuracy of their estimates. However, no reliable gender difference was observed for relative accuracy (forest plot and analyses are displayed in Appendix 3). Taken together, the present outcomes suggest that although girls/women (as compared to boys/men) are less confident in their task performance, they are equally able to discriminate between estimates that are more versus less precise.

Finally, our analyses included data from participants ranging from early childhood to adulthood, but we chose not to focus on developmental trends in our outcomes of interest. We argue that to assess developmental changes in confidence (for example), a cross-sectional or longitudinal design should be used (cf. Hutchinson et al. 2019). In particular, participants of various ages should estimate numbers within the same number-line scale (and in the datasets we analyzed, not all participants estimated numbers within the same scales). However, interpreting these data could be challenging given that the very oldest children and adults will be highly accurate in some of the smallest numerical ranges, and the very youngest children will be highly inaccurate in some of the largest numerical ranges (cf. Thompson and Opfer 2010). Alternatively, researchers could follow children from younger to older ages to assess how confidence changes over short and long periods of time. Investigating gender differences

across development is a fascinating direction for future research, but not one we can adequately address with the current dataset given we do not have enough variation in age and number-line scales to draw strong conclusions. Nevertheless, we conducted linear mixed models to statistically test for developmental effects to assess whether the magnitude of gender differences in confidence increases (or decreases) with grade and found no evidence of developmental effects in gender differences in confidence (Appendix 4).

Conclusion

The current study suggests that gender is an important source of individual differences in confidence judgments on trial-by-trial number-line estimation. In our analyses with over 700 participants who made estimates across a variety of number-line scales, we found evidence that boys/men and girls/women differ on both performance and confidence for number-line estimation. These findings advance research showing that performance is more precise for boys/men than girls/women and are the first to show that confidence judgments are higher for boys/men than girls/women, even when controlling for performance. As more researchers collect measures of confidence for number-line estimation, we hope these gender analyses could be replicated in the future. In addition, these outcomes should encourage researchers to consider analyses of gender differences for studies on math cognition and metacognition and provide pathways for future research to address our speculations about potential mechanisms underlying the gender differences observed. Research in this area also raises important questions about how gender gaps in number-line estimation performance and confidence can be narrowed and how such gaps can impact self-assessment, self-efficacy, and persistence in math and math-intensive fields.

Acknowledgements The authors gratefully acknowledge Dr. Pooja Sidney for her valuable input regarding data analysis.

Funding This work was supported in part by the Institute of Education Sciences, U.S. Department of Education award R305A160295, awarded to Clarissa A. Thompson, Kent State University.

Compliance with ethical standards

Conflict of interest We have no known conflicts of interest to disclose.

Appendix 1: Further Details on Hedges' *g* Meta-Analyses

To conduct our Hedges' *g* meta-analyses, we followed these four steps: Once an effect size (*g*) was calculated for all of the samples, we (1) estimated heterogeneity among the effect sizes, (2) estimated the population variability in effect sizes, (3) used this estimate of population variability to provide random-effects weights of sample effect sizes, and (4) used these random-effects weights to estimate a random-effects mean effect size and standard error of this estimate. In particular, heterogeneity was estimated using the *Q*-statistic (Cochran 1954), computed with the following equation:

$$Q = \Sigma\left(w_i\ (g_i - g)^2\right) \tag{A1}$$

In Eq. A1, w_i = the weight of study i ($w_i = 1/SE^2_i$, where SE_i = the standard error of the effect size estimate for study i), g_i = the effect size estimate from study i, and g_i = the mean effect size across the samples, such that Q is distributed as chi-square with k-1 degrees of freedom, where k = the number of samples. To estimate population variability in effect sizes (τ^2), the following equation was used:

$$\tau^2 = Q - (k-1) / \left\{ (\Sigma w_i) - \left[\left(\Sigma w^2{}_i \right) / (\Sigma w_i) \right] \right\} \quad \text{(A2)}$$

Random-effects weights (w^*_i) were computed as: $w^*_i = 1/(\tau^2 + SE^2_i)$, where SE_i = the standard error of the effect size of study i. Finally, the random-effects mean effect size (g) was calculated with the following equation:

$$g = \Sigma(w^* g) / \Sigma(w^*) \quad \text{(A3)}$$

The standard error of this mean effect size was computed as $SE_g = [\Sigma(w^*)]^{½}$.

Appendix 2: Analyses of Gender Differences in Performance

Hedges' *g* Meta-Analysis

A medium gender difference in precision occurred favoring boys/men, presented in Fig. 3. Averaged across the 18 effect sizes, boys/men were more precise in their number-line estimates than were girls/women, $g = .52$, 95% CI [.31, .74], $p < .001$. No significant heterogeneity was observed among the effect sizes, $Q(17) = 24.46$, $p = .11$; $I^2 = 30.50\%$. Conducting this same meta-analysis on performance using a fixed-effects model resulted in a similar overall mean effect size, $g = .48$, 95% CI [.32, .63], $p < .001$.

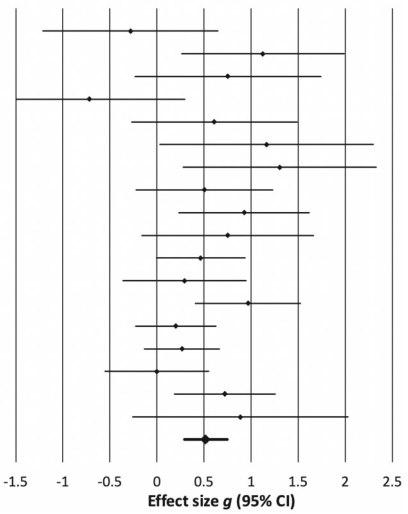
Linear Mixed Effects Model

Table 4 presents a nested linear mixed effect model predicting trial-level PAE from the participant's gender. The model replicated the findings from the Hedges' g meta-analysis on the effect of gender on estimation precision. Using the fixed-effects estimates, women's PAE is estimated to be .039 points higher than men's PAE ($p < .001$). Given the intercept of .126 ($p < .001$), the fixed-effects estimate of the average man's PAE was .126 versus .165 (.126 +

Table 4 Linear Mixed Models Predicting Trial-Level Precision (Measured as PAE)

	Model
	Coeff. (*SE*)
Fixed Effect Estimates	
Intercept (SE)	.126*** (.018)
Participant-Level (U2)	
Woman	.039*** (.007)
Random Effect Estimates	
U3 (Experiment)	.003 (.001)
U2 (Participant)	.006 (.000)
U1 (Trial)	.018 (.000)
AIC	−23,070.98
BIC	−23,031.27

Note. Trials nested within participant, which is nested within experiment. PAE = proportion absolute error. ***$p < .001$



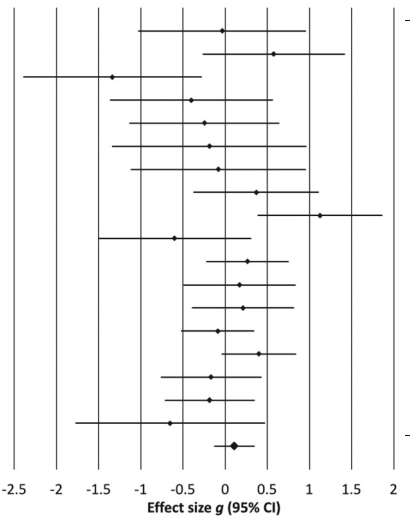
Study, Expt, and Grade	Effect size	95% CI lower	95% CI upper
Wall et al. (2016), Expt 1, Grade 1	-0.28	-1.21	0.65
Wall et al. (2016), Expt 1, Grade 2	1.12	0.26	1.99
Wall et al. (2016), Expt 1, Grade 4	0.75	-0.23	1.74
Wall et al. (2016), Expt 2, Grade 1	-0.40	-1.36	0.56
Wall et al. (2016), Expt 2, Grade 2	0.61	-0.27	1.48
Wall et al. (2016), Expt 2, Grade 4	1.17	0.03	2.30
Wall et al. (2016), Expt 3, Gorade 1	1.30	0.28	2.33
Wall et al. (2016), Expt 3, Grade 2	0.93	-0.22	2.33
Wall et al. (2016), Expt 3, Grade 4	0.93	0.23	1.62
Wall et al. (n.d.), Expt 1, Grade 2	0.75	-0.16	1.66
Fitzsimmons et al. (n.d.), Expt 1, Grade 1	0.46	-0.004	0.93
Fitzsimmons et al. (n.d.), Expt 1, Grade 2	0.29	-0.36	0.95
Fitzsimmons et al. (n.d.), Expt 2, Adults	0.97	0.41	1.52
Fitzsimmons et al. (2020), Expt 1, Adults	0.20	-0.23	0.63
Fitzsimmons et al. (2020), Expt 2, Adults	0.26	-0.13	0.66
Feltner & Thompson (n.d.), Expt 1, Grade 5	-0.002	-0.55	0.55
Fitzsimmons et al. (2019), Expt 1, Grade 4	0.72	0.19	1.26
Fitzsimmons et al. (2019), Expt 1, Grade 5	0.89	-0.26	2.03
Grand Mean	**0.52**	**0.31**	**0.74**

Fig. 3 Forest Plot of Effect Sizes for the 18 Samples Included in the Hedges' g Meta-Analysis of Performance

.039) for women. Thus, it appears that the average girl's/woman's estimate had 31% more error than the average boy's/man's estimate (.039 /.126 = .3095).

Appendix 3: Gender Differences in Relative Metacognitive Accuracy

No reliable gender difference was observed for relative accuracy, as calculated by computing an intra-individual gamma correlation between confidence judgments and performance (Fig. 4); $g = .11$, 95% CI [−.13, .35], $p = .27$. No significant heterogeneity was observed among the



Study, Expt, and Grade	Effect size	95% CI lower	95% CI upper
Wall et al. (2016), Expt 1, Grade 1	-0.03	-1.02	0.95
Wall et al. (2016), Expt 1, Grade 2	0.58	-0.26	1.41
Wall et al. (2016), Expt 1, Grade 4	-1.34	-2.39	-0.28
Wall et al. (2016), Expt 2, Grade 1	-0.40	-1.36	0.56
Wall et al. (2016), Expt 2, Grade 2	-0.25	-1.13	0.64
Wall et al. (2016), Expt 2, Grade 4	-0.19	-1.33	0.96
Wall et al. (2016), Expt 3, Grade 1	-0.08	-1.11	0.95
Wall et al. (2016), Expt 3, Grade 2	0.37	-0.37	1.10
Wall et al. (2016), Expt 3, Grade 4	1.13	0.39	1.86
Wall et al. (n.d.), Expt 1, Grade 2	-0.60	-1.50	0.30
Fitzsimmons et al. (n.d.), Expt 1, Grade 1	0.27	-0.22	0.75
Fitzsimmons et al. (n.d.), Expt 1, Grade 2	0.17	-0.49	0.83
Fitzsimmons et al. (n.d.), Expt 2, Adults	0.21	-0.39	0.81
Fitzsimmons et al. (2020), Expt 1, Adults	-0.09	-0.51	0.34
Fitzsimmons et al. (2020), Expt 2, Adults	0.40	-0.04	0.83
Feltner & Thompson (n.d.), Expt 1, Grade 5	-0.17	-0.76	0.43
Fitzsimmons et al. (2019), Expt 1, Grade 4	-0.18	-0.71	0.35
Fitzsimmons et al. (2019), Expt 1, Grade 5	-0.65	-1.77	0.47
Grand Mean	**0.11**	**-0.13**	**0.35**

Fig. 4 Forest Plot of Effect Sizes for the 18 Samples Included in the Hedges' g Meta-Analysis of Relative Metacognitive Accuracy

effect sizes, $Q(17) = 27.47$, $p = .05$, $I^2 = 38.11\%$. The fixed-effects model resulted in a similar overall mean effect size, $g = .07$, 95% CI [−.09, .23], $p = .28$.

Appendix 4: Gender Differences in Confidence (Controlling for Performance) by Grade and Number-Line Scale

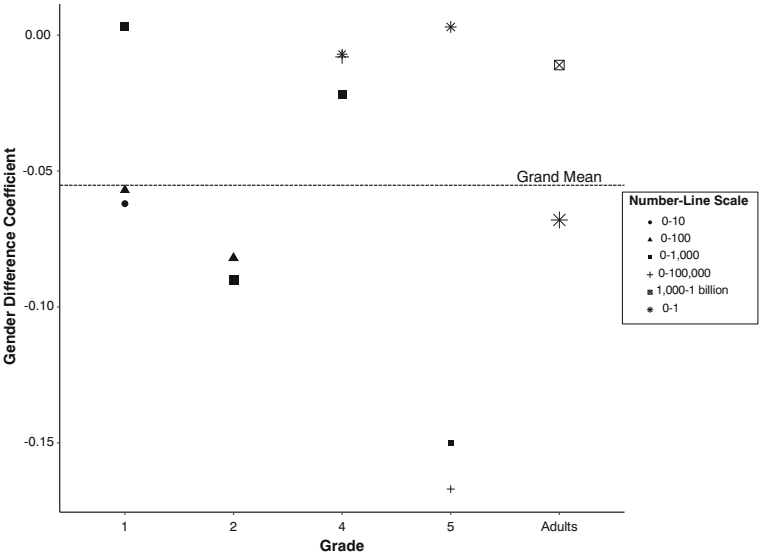


Fig. 5 Gender Differences in Confidence (Controlling for Performance) by Grade and Number-Line Scale

Note. Each point represents the effect of gender while controlling for PAE, averaged across experiments involving participants who are in the same grade and estimated within the same number-line scale. Negative values indicate boys/men are more confident than girls/women. The horizontal line represents the grand mean effect of gender ($b = -0.38$). The size of each dot is mapped to the number of observations (participants x items), with larger sizes representing more observations.

References

Data from studies preceded by an asterisk (*) were included in our analyses

American Psychological Association. (2012). Guidelines for psychological practice with lesbian, gay, and bisexual clients. *American Psychologist, 67*(1), 10–42. https://doi.org/10.1037/a0024659.

Ariel, R., Lembeck, N. A., Moffat, S., & Hertzog, C. (2018). Are there sex differences in confidence and metacognitive monitoring accuracy for everyday, academic, and psychometrically measured spatial ability? *Intelligence, 70*, 42–51. https://doi.org/10.1016/j.intell.2018.08.001.

Baer, C., & Odic, D. (2019). Certainty in numerical judgments develops independently of the approximate number system. *Cognitive Development, 52*, 100817. https://doi.org/10.1016/j.cogdev.2019.100817.

Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 17*(5), 775–785. https://doi.org/10.1111/desc.12155.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55.

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20*(5), 372–382. https://doi.org/10.1016/j.learninstruc.2009.03.002.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189.

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development, 79*(4), 1016–1031. https://doi.org/10.1111/j.1467-8624.2008.01173.x.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111. https://doi.org/10.1002/jrsm.12.

Bull, R., Cleland, A. A., & Mitchell, T. (2013). Sex differences in the spatial representation of number. *Journal of Experimental Psychology: General, 142*(1), 181–192. https://doi.org/10.1037/a0028387.

Caldera, Y. M., Huston, A. C., & O'Brien, M. (1989). Social interactions and play patterns of parents and toddlers with feminine, masculine, and neutral toys. *Child Development, 60*(1), 70–76. https://doi.org/10.2307/1131072.

Card, N, A. (2012). Applied meta-analysis for social science research. Guilford Publications

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences, 108*(8), 3157–3162. https://doi.org/10.1073/pnas.1014871108.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin, 135*(2), 218–261. https://doi.org/10.1037/a0014412.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101–129. https://doi.org/10.2307/3001666.

Colling, L. J., Szűcs, D., De Marco, D., Cipora, K., Ulrich, R., Nuerk, H. C., et al. (2020). A multilab registered replication of the attentional SNARC effect. *Advances in Methods and Practices in Psychological Science., 3*, 143–162. https://doi.org/10.1177/2515245920903079.

Cooke-Simpson, A., & Voyer, D. (2007). Confidence and gender differences on the mental rotations test. *Learning and Individual Differences, 17*(2), 181–186. https://doi.org/10.1016/j.lindif.2007.03.009.

Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology, 106*(6), 1691–1730. https://doi.org/10.1086/321299.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*, 371–396. https://doi.org/10.1037/0096-3445.122.3.371.

Doyle, R. A., Voyer, D., & Cherney, I. D. (2012). The relation between childhood spatial activities and spatial abilities in adulthood. *Journal of Applied Developmental Psychology, 33*(2), 112–120. https://doi.org/10.1016/j.appdev.2012.01.002.

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *Oxford library of psychology. The Oxford handbook of metamemory (p. 23–37)*. Oxford University Press.

Dwyer, C. A., & Johnson, L. M. (1997). Grades, accomplishments, and correlates. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment (p. 127–156)*. Lawrence Erlbaum Associates Publishers.

Ellis, J., Fosdick, B. K., & Rasmussen, C. (2016). Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. *PloS one, 11*(7), e0157447. https://doi.org/10.1371/journal.pone.0157447.

Estes, Z., & Felker, S. (2012). Confidence mediates the sex difference in mental rotation performance. *Archives of Sexual Behavior, 41*(3), 557–570. https://doi.org/10.1007/s10508-011-9875-5.

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology, 123*, 53–72. https://doi.org/10.1016/j.jecp.2014.01.013.

*Feltner, A. & Thompson, C, A. (n.d.). Playing to learn: Helping children learn fraction magnitudes through board games.

*Fitzsimmons, C., Rivers, M, L., Sidney, P, G., Dunlosky, J., & Thompson, C, A. (2019). What cues do children use when judging their confidence in fraction estimation performance? Confidence judgments relate more strongly to familiarity than performance. Poster presented at the Cognitive Development Society Biennial Conference, Louisville, KY.

*Fitzsimmons, C., Thompson, C, A., & Sidney, P, G. (2020). Confident or familiar? The role of familiarity ratings in adults' confidence judgements when estimating fraction magnitudes. *Metacognition and Learning.* https://doi.org/10.1007/s11409-020-09225-9, 15, 215, 231.

*Fitzsimmons, C., Morehead, K., Thompson, C, A., Buerke, M., & Dunlosky, J. (invited revision). Does studying worked examples improve numerical magnitude estimation?

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., & Bryant, J. D. (2010). The contributions of numerosity and domain-general abilities to school readiness. *Child Development, 81*, 1520–1533. https://doi.org/10.1111/j.1467-8624.2010.01489.x.

Ganley, C. M., & Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. *Learning and Individual Differences, 47*, 182–193. https://doi.org/10.1016/j.lindif.2016.01.002.

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology, 47*(6), 1539–1552. https://doi.org/10.1037/a0025510.

Goldstein, J. M., Seidman, L. J., Horton, N. J., Makris, N., Kennedy, D. N., Caviness Jr., V. S., Faraone, S. V., & Tsuang, M. T. (2001). Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging. *Cerebral Cortex, 11*(6), 490–497. https://doi.org/10.1093/cercor/11.6.490.

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*(5), 1229–1241. https://doi.org/10.1037/a0027433.

Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science, 13*(4), 135–139. https://doi.org/10.1111/j.0963-7214.2004.00292.x.

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93*(2), 388–395. https://doi.org/10.1037/0033-2909.93.2.388.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486.

Hembacher, E., & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768–1776. https://doi.org/10.1177/0956797614542273.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557.

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience, 6*(6), 435–448. https://doi.org/10.1038/nrn1684.

Hutchinson, J. E., Lyons, I. M., & Ansari, D. (2019). More similar than different: Gender differences in children's basic numerical skills are the exception not the rule. *Child Development, 90*(1), e66–e79. https://doi.org/10.1111/cdev.13044.

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: A meta-analysis. *Psychology of Women Quarterly, 14*(3), 299–324. https://doi.org/10.1111/j.1471-6402.1990.tb00022.x.

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology, 42*(1), 11–26. https://doi.org/10.1037/0012-1649.42.1.11.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*(2), 198–214. https://doi.org/10.1037/0033-2909.105.2.198.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349.

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition, 19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010.

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology (p. 483–502)*. The Guilford Press.

Lauer, J. E., Yhang, E., & Lourenco, S. F. (2019). The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin, 145*(6), 537–565. https://doi.org/10.1037/bul0000191.

Lawson, G. M., Hook, C. J., & Farah, M. J. (2018). A meta-analysis of the relationship between socioeconomic status and executive function performance among children. *Developmental Science, 21*(2), e12529. https://doi.org/10.1111/desc.12529.

LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development, 81*(6), 1753–1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x.

Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., & Ratliff, K. (2016). Sex differences in spatial cognition: Advancing the conversation. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(2), 127–155. https://doi.org/10.1002/wcs.1380.

Marshman, E. M., Kalender, Z. Y., Nokes-Malach, T., Schunn, C., & Singh, C. (2018). Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Physical Review Physics Education Research, 14*(2), 020123. https://doi.org/10.1103/PhysRevPhysEducRes.14.020123.

McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology, 27*(5), 486–493. https://doi.org/10.1016/j.appdev.2006.06.003.

Moè, A., & Pazzaglia, F. (2006). Following the instructions!: Effects of gender beliefs in mental rotation. *Learning and Individual Differences, 16*(4), 369–377. https://doi.org/10.1016/j.lindif.2007.01.002.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109.

Nelson, J. A. (2014). The power of stereotyping and confirmation bias to overwhelm accurate assessment: The case of economics, gender, and risk aversion. *Journal of Economic Methodology, 21*(3), 211–231. https://doi.org/10.1080/1350178X.2014.939691.

Nelson, L. J., & Fyfe, E. R. (2019). Metacognitive monitoring and help-seeking decisions on mathematical equivalence problems. *Metacognition and Learning, 14*, 167–187. https://doi.org/10.1007/s11409-019-09203-w.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55*(3), 169–195. https://doi.org/10.1016/j.cogpsych.2006.09.002.

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development, 79*(3), 788–804. https://doi.org/10.1111/j.1467-8624.2008.01158.x.

Opfer, J. E., Thompson, C. A., & Furlong, E. (2010). Early development of spatial-numeric associations: Evidence from spatial and quantitative performance of preschoolers. *Developmental Science, 13*, 761–771. https://doi.org/10.1111/j.1467-7687.2009.00934.x.

Pruden, S. M., & Levine, S. C. (2017). Parents' spatial language mediates a sex difference in preschoolers' spatial language use. *Psychological Science, 28*(11), 1583–1596. https://doi.org/10.1177/0956797617711968.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russel Sage Foundation.

Reinert, R. M., Huber, S., Nuerk, H. C., & Moeller, K. (2017). Sex differences in number line estimation: The role of numerical estimation. *British Journal of Psychology, 108*(2), 334–350. https://doi.org/10.1111/bjop.12203.

Rosenthal, R., Rosnow, R, L., & Rubin, D, B. (2000). Contrasts and effect sizes in behavioral research: A correlational approach. Cambridge University Press.

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development, 89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068.

Sidney, P, G., Thompson, C, A., Fitzsimmons, C., & Taber, J, M. (2019). Children's and adults' math attitudes are differentiated by number type. The Journal of Experimental Education, 1-32. https://doi.org/10.1080/00220973.2019.1653815

Siegler, R. S. (2016). Magnitude knowledge: The common core of numerical development. *Developmental Science, 19*(3), 341–361. https://doi.org/10.1111/desc.12395.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*(3), 237–250. https://doi.org/10.1111/1467-9280.02438.

Siegler, R. S., & Thompson, C. A. (2014). Numerical landmarks are useful—Except when they're not. *Journal of Experimental Child Psychology, 120*, 39–58. https://doi.org/10.1016/j.jecp.2013.11.014.

Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain, and Education, 3*(3), 143–150. https://doi.org/10.1111/j.1751-228X.2009.01064.x.

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001.

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, I., & Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23*(7), 691–697. https://doi.org/10.1177/0956797612440101.

Syzmanowicz, A., & Furnham, A. (2011). Gender differences in self-estimates of general, mathematical, spatial and verbal intelligence: Four meta analyses. *Learning and Individual Differences, 21*(5), 493–504. https://doi.org/10.1016/j.lindif.2011.07.001.

Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and sex differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology, 101*(1), 20–51. https://doi.org/10.1016/j.jecp.2008.02.003.

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development, 81*(6), 1768–1786. https://doi.org/10.1111/j.1467-8624.2010.01509.x.

Tosto, M. G., Garon-Carrier, G., Gross, S., Petrill, S. A., Malykh, S., Malki, K., Hart, S. A., Thompson, L., Karadaghi, R. L., Yakovlev, N., Tikhomirova, T., Opfer, J. E., Mazzocco, M. M. M., Dionne, G., Brendgen, M., Vitaro, F., Tremblay, R. E., Boivin, M., & Kovas, Y. (2018). The nature of the association between number line and mathematical performance: An international twin study. *British Journal of Educational Psychology, 89*, 787–803. https://doi.org/10.1111/bjep.12259.

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition, 46*(4), 507–519. https://doi.org/10.3758/s13421-017-0780-6.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250–270. https://doi.org/10.1037/0033-2909.117.2.250.

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin & Review, 24*(2), 307–334. https://doi.org/10.3758/s13423-016-1085-7.

*Wall, J, L., Thompson, C, A., Dunlosky, J., & Merriman, W, E. (2016). Children can accurately monitor and control their number-line estimation performance. Developmental Psychology, 52(10), 1493–1502. https://doi.org/10.1037/dev0000180.

*Wall, J, L., Feltner, A., Merriman, W, E., Dunlosky, J., & Thompson, C, A. (n.d.). Which numbers do children choose to re-estimate when given the opportunity to control their performance?