Accepted for publication in Journal of Experimental Child Psychology

doi 10.1016/j.jecp.2019.01.012

More than number sense: The additional role of executive functions and metacognition in arithmetic

Elien Bellon[1a], Wim Fias[2] and Bert De Smedt[1]

[1] Psychology and Educational Sciences, Parenting and Special Education Research Group, KU Leuven, Leopold Vanderkelenstraat 32, 3000 Leuven, Belgium

[2] Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Gent, Belgium

[a] Corresponding author: elien.bellon@kuleuven.be

RESEARCH HIGHLIGHTS

- Simultaneous investigation of domain-specific and domain-general cognitive factors
- Unique roles of updating, metacognition, & magnitude skills in arithmetic
- On-task metacognitive monitoring is a very powerful unique predictor of arithmetic
- Symbolic numerical magnitude processing remains a unique predictor of arithmetic

Abstract

Arithmetic is a major building block for children's development of more complex mathematical abilities. Knowing which cognitive factors underlie individual differences in arithmetic is key to gain further insight into children's mathematical development. The current study investigated the role of executive functions and metacognition (domain-general cognitive factors) as well as symbolic numerical magnitude processing (domain-specific cognitive factor) in arithmetic, enabling the investigation of their unique contribution in addition to each other. Participants were 127 typically developing second graders (7- to 8-year-olds). Our within-subject design took into account different components of executive functions (i.e., inhibition, shifting and updating), different aspects of metacognitive skills (i.e., task-specific and general metacognition), and different levels of experience in arithmetic, namely addition – in which second graders had extensive experience-, and multiplication – for which second graders were still in the learning phase. This study reveals that both updating and metacognitive monitoring are important unique predictors of arithmetic, in addition to each other and in addition to symbolic numerical magnitude processing. Our results point to a strong and unique role of task-specific metacognitive monitoring skills. These individual differences in noticing your own errors might help you to learn from your mistakes.

Keywords: executive functions; metacognition; numerical magnitude processing; arithmetic

Introduction

While waiters know within seconds how much the different drinks you ordered cost together, others shudder when only thinking about math class. Experience with mathematics could be a part of the explanation for these huge differences in math skills in adults, yet these differences already appear early in development (e.g., Duncan et al., 2007). An extensive body of literature demonstrates that there are large individual differences in the way children acquire mathematical abilities (e.g., Berch, Geary, & Mann Koepke, 2016). When looking for cognitive explanations for these differences, research has investigated both domain-general cognitive factors (i.e., factors relevant for learning various academic skills) and domain-specific cognitive factors (i.e., factors relevant for mathematics learning in specific) (e.g., Geary & Moore, 2016; Vanbinst & De Smedt, 2016), yet these types of factors have been investigated in relative isolation from each other. Over the last decade, research has mainly focused on one domain-specific cognitive skill, numerical magnitude processing (i.e., *children's elementary intuitions about quantity and the ability to understand the meaning of number*s), as a core factor of individual differences in mathematical abilities (e.g., De Smedt, Noël, Gilmore, & Ansari, 2013, for a review; Schneider et al., 2017, for a meta-analysis). By narrowing its scope in this way, this research ignores other critical, particularly domain-general, cognitive factors that might play a role in (a)typical mathematical development (see Fias, Menon, & Szucs, 2013). Schneider and colleagues (2017) revealed in their meta-analysis that the overall correlation between numerical magnitude processing and mathematical competence was only r = .278, suggesting numerical magnitude skills explain a significant but only small part of the variance in mathematical abilities. Consequently, other important factors that drive these individual differences need to be considered.

Some potentially important factors have been identified, such as working memory (e.g., Peng, Namkung, Barnes, & Sun, 2016) and other executive functions (e.g., Bull & Lee, 2014). However, the precise roles of these factors in mathematics in addition to numerical magnitude processing remain unclear. Executive functions refer to a family of top-down mental processes needed when you have to

concentrate and pay attention (Diamond, 2013), skills required to monitor and control thought and action, processes that allow us to respond flexibly to our environment and engage in deliberate, goal-directed, thought and action (Cragg & Gilmore, 2014). Executive functions mainly consist of the processes of inhibition, shifting and updating (e.g., Miyake et al., 2000). Inhibition refers to one's ability to control his/her attention, behaviour, and thoughts to override a strong internal predisposition or external lure and instead do what is more appropriate or needed (Diamond, 2013). Two types of inhibition can be distinguished, namely response inhibition (i.e., control over one's behaviour and control over one's emotions in the service of controlling one's behaviour, to resist temptations and not acting impulsively) and interference control (i.e., the ability to selectively attend, focusing on what we choose and suppressing attention to other stimuli). Shifting is defined as the disengagement of an irrelevant task set or strategy, and the subsequent initiation of a new, more appropriate set (van der Sluis, de Jong, & van der Leij, 2007). Updating involves holding information in working memory and flexibly manipulating it (Baddeley & Hitch, 1994).

While existing literature has provided some understanding about the association between executive functions and mathematics (Cragg & Gilmore, 2014), there is insufficient insight into the specificity of these associations. Some studies found a significant association between executive functions and mathematics, others did not (e.g., Bull & Lee, 2014; Cragg & Gilmore, 2014; Friso-Van Den Bos, Van Der Ven, Kroesbergen, & Van Luit, 2013).

This inconclusiveness might be explained by four limitations in the existing literature. Firstly, mathematical ability consists of various different abilities (e.g., arithmetic, problem solving, geometry) which could all be differently related to certain cognitive factors (e.g., Cragg & Gilmore, 2014, for a critical discussion). Consequently, focussing on different math subcomponents or tasks could be the reason for these discrepancies in existing research. For example, Gilmore and colleagues (2015) found different associations between inhibition and specific components of mathematics. Therefore, investigating specific mathematical skills is essential when investigating which cognitive factors are

predicting its development. The current study focused on arithmetic, a major building block for children's development in more complex mathematical abilities (Kilpatrick, Swafford, & Findell, 2001). The association between executive functions and arithmetic is theoretically appealing. For example, cognitive inhibition (i.e., interference control) might play a role in arithmetic, as arithmetic facts are stored in an associative network in semantic memory (e.g., Verguts & Fias, 2005) and thus are particularly prone to interference because of the number of features they share (De Visscher & Noël, 2014a). Hence, during arithmetic, incorrect but competing answers have to be inhibited. On the other hand, good executive functioning might help to suppress inefficient arithmetic strategies in favour of more efficient ones (e.g., using a decomposition of operands strategy instead of finger counting), or might help to keep intermediate solutions in mind and manipulating them while calculating (i.e., updating).

Secondly, the discrepancy between results on the relation between executive functions and mathematics might be due to the multi-componential nature of executive functions. Executive functioning includes different dimensions, yet many studies take only one subcomponent into account. Different components of executive functions are differently related to mathematics (Bull & Lee, 2014; Cragg, Keeble, Richardson, Roome, & Gilmore, 2017). As there is both unity and diversity in the executive functioning components (Friedman & Miyake, 2017; Lee, Bull, & Ho, 2013), it is important to investigate them simultaneously. In this study, we included all three well-known components of executive functions, namely inhibition, shifting and updating.

Thirdly, the inconclusiveness might be due to the fact that relevant moderating variables have been overlooked. One such variable that has often been neglected in theoretical models of executive functions (see Desender, Van Opstal, & Van den Bussche, 2014; for a critical discussion), but has received a lot of attention in more general educational research (e.g., Schneider & Artelt, 2010), is metacognition. Metacognition was first introduced by Flavell (1979) as a broader concept that encompasses monitoring and regulation of cognitive performance. It involves the ability to assess

one's own cognitive knowledge and ability (Vo, Li, Kornell, Pouget, & Cantlon, 2014) and how people monitor and control their cognition on-task (Bryce, Whitebread, & Szűcs, 2015). An important distinction is made between metacognitive monitoring and metacognitive control (Flavell, 1979; Nelson & Narens, 1990). The former is *the meta-level that is informed by the object-level* and the latter the m*eta-level that modifies the object-level*, with the object-level being the cognitive process or task at hand. Metacognition can be measured by means of general metacognitive knowledge/skills measures (e.g., Haberkorn, Lockl, Pohl, Ebert, & Weinert, 2014), or by asking participants to indicate their confidence on a task-specific level (e.g., Rinne & Mazzocco, 2014).

The association between metacognition and general academic performance has been extensively studied. Vo and colleagues (2014) found that children's metacognitive ability in the numerical domain predicted their general school-based mathematics knowledge, and suggested that children's metacognition is a domain-dependent cognitive ability in children. Freeman, Karayanidis, and Chalmers (2017) found that calibration of confidence in a working memory task was related to academic achievement. Some studies have specifically focussed on the association between metacognition and arithmetic. On a theoretical level, Shrager and Siegler (1998) identified a metacognitive system in their model of children's strategy choices and discoveries in arithmetic (i.e., SCADS model), in which a confidence criterion (i.e., a randomly varying threshold for stating a retrieved answer) was one of the most important probabilistic components. Rinne and Mazzocco (2014) more recently showed that, in children in late elementary school, on-task metacognitive judgments were strongly related to concurrent mental arithmetic, and that early metacognitive monitoring ability predicted subsequent growth in mental arithmetic performance. However, it remains to be determined whether these associations remain when the different aspects of executive functions as well as numerical magnitude processing are included, and whether the same associations can be observed in younger children.

Importantly, metacognition and executive functions are closely linked. Both are higher-order factors related to the regulation of behaviour, they share theoretical features (e.g., controlled processing), undergo similar developmental trajectories, and are associated with comparable brain regions (Roebers & Feurer, 2015). Metacognition plays a critical role in executive functioning processes, because subjective experiences (i.e., metacognition) allow for top-down control of behaviour (Desender et al., 2014). Indeed, effective executive functioning requires an accurate determination of when and which type of control is needed. On the other hand, Bryce and colleagues (2015) suggest that executive functions might be *necessary but not sufficient* antecedents to metacognitive skills. Lyons and Zelazo (2011) characterized metacognitive monitoring as a reflective process. This reflective process is explicitly conceptualized as an integral part of metacognition (Nelson & Narens, 1990), but is only implicitly assumed to take place in executive functions, for which post error slowing can be an indicator. Since executive functions and metacognition skills are closely related (see Roebers, 2017, for a review), the lack of inclusion of metacognition in existing studies on the association between mathematics and executive functions may have led to inconclusive results.

Lastly, the inconclusiveness in the current literature on the association between executive functions and mathematics might be due to the age of the participants. As arithmetic (Vanbinst, Ceulemans, Ghesquière, & De Smedt, 2015), executive functions (Carlson, Zelazo, & Faja, 2013) and metacognition (Schneider, 2010) are still in the midst of development during primary school, different associations might be present at different ages. Executive functioning develops substantially during primary school (e.g., Diamond, 2013), with increases in inhibition, shifting and updating skills allowing for better executive functioning along with increasing automaticity and efficiency (Carlson et al., 2013). Executive functioning depends upon neural networks involving the prefrontal cortex, which continue to improve into early adulthood. Metacognition also relies on the prefrontal cortex (Fleming & Dolan, 2014), and continues to develop through primary school (e.g., Schneider, 2015; Schneider & Lockl, 2008).

Arithmetic skills also show a large development in primary school and a gradual change from procedural strategies to retrieval of arithmetic facts is observed (Siegler, 1996). This might lead to different associations with executive functions and metacognition over time. For example, the need for inhibition skills might be particularly high when children have already learned (some) arithmetic facts, because similarity between these facts provokes interference (De Visscher & Noël, 2014b). Consequently, the association between executive functions and arithmetic could change across development. One prediction might be that inhibition skills have a time-limited role in arithmetic: At the early stages of the development from procedural strategies to full automatization of arithmetic fact retrieval, more interference control skills will be needed; at the end of development, the need for control would be lower as a semantic network is automatically activated (see Moors & De Houwer, 2006; Siegler, 1996; for a theoretical discussion).

As these three cognitive factors (i.e., arithmetic, executive functions and metacognition) develop dramatically throughout primary school, the association of executive functions and metacognition with arithmetic might be variable across development. Hence, the level of experience in arithmetic is important to take into account when examining the association between executive functions, metacognition and arithmetic. Accordingly, in this study in second graders, we included an arithmetic operation in which they had extensive experience (i.e., addition), and retrieval of single-digit arithmetic items is very high (De Smedt, 2016), as well as an operation for which participants were in the learning phase (i.e., multiplication) and there is no full automatization of fact retrieval present yet.

The Present Study

Little attention has been paid to investigating the joint effects of domain-specific and domain-general correlates of arithmetic. Additionally, it is not clear how these factors are associated in young children (i.e., early elementary school), and consequently, how these factors play a role in function of lesser versus greater experience in arithmetic. To tackle this important gap in the literature, the present study in typically developing second graders investigated executive functions – by using a more

fine-grained operationalization including inhibition, shifting and updating through the use of tasks that tap into those different aspects – and metacognition – both general metacognitive knowledge and on-task metacognitive monitoring – as domain-general cognitive factors, and symbolic numerical magnitude processing as a domain-specific cognitive factor, and examined their unique contributions to arithmetic (i.e., addition and multiplication) using a within-subject design. We specifically recruited second graders to examine these associations in arithmetic operations in which they had extensive experience (addition) and in which they were still in the learning phase (multiplication).

We predicted that executive functioning would be associated with arithmetic (i.e., better executive functioning is associated with better, faster and more accurate, arithmetic performance). Specifically, we expected the strongest association between inhibition skills and the arithmetic task in which participants had the most experience (i.e., addition). Secondly, we predicted that metacognition would be an important predictor of arithmetic (i.e., better metacognitive skills are associated with better arithmetic performance). Based on the suggestion of domain-dependent metacognitive abilities, we predicted that this association would be the strongest for the on-task metacognitive monitoring. Thirdly, we predicted that executive functions, metacognition, and symbolic numerical magnitude processing would each explain unique variance in individual differences in arithmetic.

Method

Participants

Participants were 127 Flemish second-graders (64 girls, 63 boys; M_{age} = 7 years, 11 months; *SD* = 4 months; [6y10m-8y7m]). They were all typically developing children, who had no diagnosis of a developmental disorder, did not repeat a grade and had a dominantly middle- to high socio-economic background. For every participant, written informed parental consent was obtained. The study was approved by the Social and Societal Ethics Committee of the KU Leuven.

Materials

Materials consisted of standardized tests, paper-and-pencil tasks, and computer tasks designed with E-Prime 2.0 (Schneider, Eschmann, & Zuccolotto, 2002) and Affect 4.0 (Spruyt, Clarysse, Vansteenwegen, Baeyens, & Hermans, 2010).

Arithmetic.

Arithmetic was assessed with two single-digit computerized production tasks, i.e., addition and multiplication. Additions and multiplications were presented in separate tasks. Arithmetic problems with 0 and 1 as one of the operands were excluded. Both the addition and multiplication items comprised all combinations of the numbers 2 to 9 as operands, and all commutative pairs (e.g., 3 + 9 and 9 + 3) were presented, yielding 64 additions and 64 multiplications. After fixation, stimuli occurred for 2000 ms in white on a black background (Arial font, 72-point size). Afterwards, a black screen appeared until response. Children were asked to answer verbally and as quickly and accurately as possible. This specific protocol was used to foster retrieval, and increase the likelihood of making errors (see below). The experimenter registered the response time and answer via the computer. For each task, stimuli were pseudo randomly divided into two blocks (i.e., one of each commutative pair in each block). During the second block of each arithmetic task, a specific metacognitive monitoring measure was added to the task (see below). Each task began with six practice items to familiarize the children with the task requirements. Performance measures were the average response time and accuracy of the answers, which were calculated for each operation separately.

Executive functions.

Executive functioning was measured with inhibition, shifting and updating tasks. None of the executive functioning tasks included numerical stimuli, which allowed us to investigate executive functions without numerical processing confounds.

Inhibition. We used a classic arrow **Flanker task** (Eriksen & Eriksen, 1974; Huizinga, Dolan, & van der Molen, 2006): a speeded choice reaction time task where participants had to respond to target stimuli (i.e., a left or right pointing arrow presented at the centre of the screen), flanked by distractors

(i.e., two arrows on each side of the target). These four distracter arrows either pointed in the same direction as the target, and thus indicated the same response (i.e., → → → → → or ← ← ← ← ←; congruent condition), or the opposite direction, and thus indicated the response that is incompatible with the target (i.e., → → ← → → or ← ← → ← ←; incongruent condition). Children had to indicate the direction of the middle arrow by pressing the corresponding key (i.e., left/right key) while suppressing the direction of the distractors. In total 40 test items were administered, 50% of the items were incongruent, and the direction of the middle arrow was balanced. After fixation, stimuli appeared for 2500 ms, followed by a black screen that remained visible for 1000 ms. Before the task started, 12 trials were presented – used as a baseline condition for the analyses of this task – on which only one arrow was presented in the centre of the screen (neutral condition). Children had to indicate the direction of the arrow. Response time and accuracy of the answers were registered by the computer.

We also used an **animal Stroop task** (based on Szucs, Devine, Soltesz, Nobes, & Gabriel, 2013). The task consisted of comparing two simultaneously presented images of coloured animals arranged on either side of the centre of the screen. One animal was selected from a set of large animals (e.g., giraffe, moose), the other was selected from a set of small animals (e.g., rabbit, frog). Participants had to select the animal which was larger in real life by pressing the corresponding key (i.e., left/right key). They were required to ignore the size of the images on the screen and to respond based on the animal size in real life only. To ensure all children had the real-world knowledge needed for this task, participants were shown the animal images one-by-one in a standard size prior to the task and were asked whether the animal was large or small in real life. After this, children were presented with 20 trials – used as a baseline condition for the analyses of this task – in which the animals occupied the same area on the screen (neutral condition). On each test trial one animal image was presented with an area on the screen four times larger than the other image. This yielded two conditions (Figure 1): a congruent condition (i.e., larger animal in real life = larger image on the screen) and an incongruent condition (i.e., larger animal in real life = smaller image on the screen). The position of the larger animal (both in real life and the size of the image) was balanced. In total 40 test trials were administered, 50%

of the items were incongruent, and items were presented in a random order. After fixation, stimuli appeared for 2000 ms, followed by a black screen which remained visible for 1000 ms. Response time and accuracy of the answers were registered by the computer.

As both accuracy and response time constitute essential parts of inhibition, we calculated inverse efficiency scores (i.e., response time divided by accuracy) for both inhibition tasks. These scores were used as performance measure for both tasks (see also Supplementary material). The use of two inhibition tasks enabled us to differentiate between the information that needs to be inhibited in the two tasks, i.e., visually, stimulus driven information in the Flanker task vs. linked to acquired knowledge in the animal Stroop task.

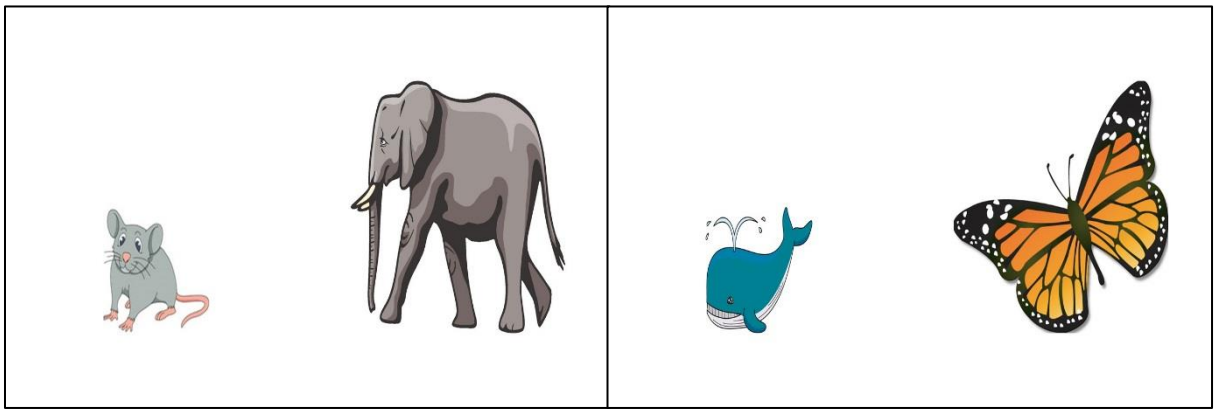

Figure 1. Example stimuli for the animal Stroop task. Left: congruent condition; Right: incongruent condition.

Shifting. To measure shifting skills, we used the Wisconsin Card Sorting Task (WCST; Grant & Berg, 1948). In this sorting task, participants had to determine how to sort cards on the basis of unspecified categories (i.e., shape and colour). We only used two sorting rules from the original WCST to minimize working memory demands (Watson, Azizian, & Squires, 2006). On each trial, three cards were simultaneously presented on a computer screen, one at the top and two at the bottom (i.e., left and right). Each card consisted of a figure with a specific shape and colour. Children had to indicate which of the two bottom cards matched the top card. The bottom cards were either identical to the

top card in shape (but different in colour) or in colour (but different in shape). On two successive items, the cards had different colours and different shapes. Children were given no explicit instructions about the sorting rules. These had to be inferred based on the feedback which was given after every item. Without notice, the sorting rule changed after a variable number (seven, eight or nine) of consecutive correct responses. After this switch item children had to disengage from the previous sorting rule and discover and/or apply the other sorting rule. The items were randomly divided in five runs, each consisting of five blocks. In each block a particular sorting rule had to be applied. A new block began after the sorting rule changed. In total, the sorting rule changed four times in each run, and thus 20 times during the whole task. When a child failed to reach seven, eight or nine consecutive correct responses within 50 trials, the task was discontinued. To familiarize the children with the task, one practice block was presented in which cards had to be sorted by shape; after eight consecutive correct responses, this block was completed. Children had to answer by pressing the corresponding response key (i.e., left/right key). The item remained visible until response. The performance measure was the average number of items a child needed to switch between rules (i.e., total number of items needed to switch divided by number of blocks completed).

Updating skills. Updating was assessed by means of a standard 2-back task (Pelegrina et al., 2015). In this continuous recognition task, a sequence of items is shown. For each item, participants had to indicate whether the presented stimulus was identical to the stimulus presented two trials back by pressing a green or red key, hereby answering yes or no respectively. Hence, children had to store two elements in their memory and update these elements, given that each time, they needed to eliminate the previous item, add the new one and maintain the presentation order of the items. Items were coloured images (e.g., rocket, sweater) presented one by one in the centre of a white screen. In total 40 items, divided in two blocks, each consisting of 20 trials, were presented. In each block, 30% of the trials were target trials (i.e., correct answer = yes). The first three trials of each block were always non-target items. After fixation, the stimulus appeared for 3000 ms, followed by a black screen that

remained visible for 1000 ms. To familiarize the children with the task, one practice block (20 items) was presented. The performance measure was the accuracy of the answer.

Metacognition.

Two aspects of metacognition, namely general metacognitive knowledge, and task-specific metacognitive judgments incorporated in the arithmetic task, were included in this study to explore both domain-general and domain-specific aspects of metacognitive skills.

Both dimensions on which metacognitive judgments can vary, that is time of judgement (i.e., pre-diction versus post-diction) and granularity of judgements (i.e., local judgments specific to one item versus global holistic judgments across multiple items) were included (Pieschl, 2009). On the one hand, one local metacognitive question measuring calibration of confidence was asked after every arithmetic item in the second block of the arithmetic task. On the other hand, two global metacognitive questions were asked, one before the start of the arithmetic task and one after the entire task was finished.

By selecting confidence judgements as our task-specific metacognitive measure and not asking participants to exhibit control actions (e.g., make modifications to or correct their answers), we focused on the monitoring aspect of metacognition rather than on its control mechanism (Nelson & Narens, 1990).

General metacognitive knowledge. To measure metacognitive abilities independently of arithmetic, we used a general metacognitive questionnaire (Haberkorn, Lockl, Pohl, Ebert, & Weinert, 2014). In this questionnaire, 15 situations involving mental performance (e.g., *Which strategy do you think is better to make sure you won't forget to take your skates to school the next day?*) were described and three possible answers (e.g., *a. Write a note on a piece of paper; b. Think strongly about the skates; c. Both proposed strategies are equally good/bad*) were presented. The researcher read the situations and the corresponding options aloud one by one. Children were given a response form with

pictures of the three possible answers (see Figure 2), so that they could follow each item and indicate their answer. The performance measure was the number of correct answers.

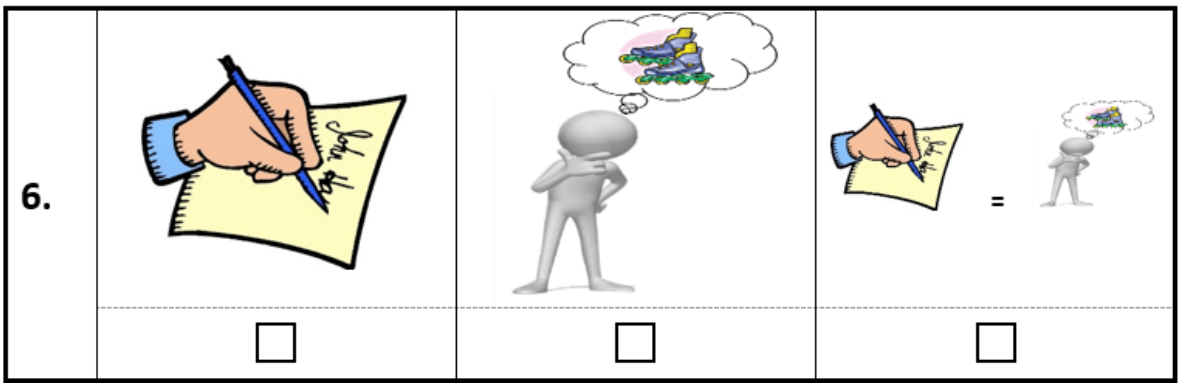

Figure 2. Example of the response form of the general metacognitive knowledge questionnaire.

Local, on-task metacognitive monitoring. Calibration of confidence, i.e., the alignment between one's confidence in solving a problem and the accuracy of the answer, was measured as in Rinne and Mazzocco (2014), by asking children on a trial-by-trial basis to report their confidence in the accuracy of their answer during the arithmetic tasks. These data were collected for every arithmetic item in the second block of both arithmetic tasks (n = 32 in each task). A specific protocol was used in the arithmetic tasks to foster retrieval and to increase the likelihood of making errors, and thus to maximize the variability in this metacognitive monitoring measure. After giving their answer to the arithmetic problem, children had to indicate how confident they were that their answer was correct (i.e., *Correct*, *I am not sure*, *Incorrect*). Calibration scores were the alignment between children's confidence rating and the accuracy of their answer (i.e., correct arithmetic answers yielded a score of 2 if children said they were *Correct*, 1 if they said *I am not sure*, and 0 if they said *Incorrect*; this scale was reversed when the arithmetic answer was incorrect). The calibration score per child was the mean of all calibration scores (i.e., calibration score per item; n = 32) and was calculated for each arithmetic task separately. The higher the calibration scores, the better the on-task metacognitive monitoring skills. To familiarize the children with the task, six practice items were presented. We used emoticons

in combination with the options (e.g., ☺ and *Correct*) to make the task more attractive and feasible for children.

Global domain-specific metacognitive monitoring. Two global metacognitive questions were asked in both arithmetic tasks. Before the task started, children had to report how they thought they would perform on the task (i.e., prospective question). After the task was finished, they had to report how they thought they had performed on the task (i.e., retrospective question). During data-analyses, we observed that the scoring of these items was problematic and we decided to remove these two questions (see Supplementary materials section).

Numerical magnitude processing.

A numerical magnitude comparison task was used to assess children's numerical magnitude processing skills. In view of the meta-analysis by Schneider and colleagues (2017), which showed that symbolic measures were stronger predictors of mathematical performance than non-symbolic measures, we administered a symbolic version of the task using Arabic digits as stimuli. This task consisted of comparing two simultaneously presented numerical magnitudes (i.e., Arabic digits) on either side of the centre of the screen. Participants had to select the numerically larger numerical magnitude by pressing the corresponding response key (i.e., left/right key). Stimuli comprised all combinations of 1 to 9, yielding 72 trials. The trials were randomly divided into two blocks. After fixation, stimuli appeared until response. The position of the largest digit was balanced. To familiarize the children with the task requirements, three practice trials were presented. The performance measure was average response time of correct responses.

Control variables.

Intellectual ability was assessed by means of the Raven's Standard Progressive Matrices (Raven, Court, & Raven, 1992). Children were given 60 multiple-choice items in which they had to complete a pattern. The performance measure was the number of correctly solved patterns.

A ***motor speed task*** was included as a control for children's response speed on the keyboard (De Smedt & Boets, 2010). Two shapes were simultaneously presented on either side of the screen and children had to indicate which of the two shapes was filled by pressing the corresponding key (i.e., left/right key). All shapes were similar in size and each shape occurred four times filled and four times non-filled, yielding 20 trials. The position of the filled shape was balanced. After fixation, stimuli appeared until response. Three practice trials were included to familiarize the children with the task. The performance measure was average response time of correct responses.

Procedure

All participants were tested at their own school during regular school hours. They all completed three sessions: an individual session including the arithmetic tasks, a session in small groups of five children including the motor speed task, the symbolic numerical magnitude processing task, and the executive functioning tasks, and a group-administered session including the metacognitive questionnaire and intellectual ability task. The sessions took about 40, 45 and 60 minutes respectively. All children completed the tasks in the abovementioned order.

Analyses

We ran frequentist analyses using both uni- and multivariate techniques, as well as Bayesian analyses. Frequentist analyses allowed us to explore our data by means of a well-known method to gage statistical support for the hypotheses of interest. However, this *p*-value null hypothesis significance testing has a number of statistical limitations (Andraszewicz et al., 2015). For example, *p*-values cannot quantify evidence in favour of a null hypothesis, they only signal the extremeness of the data under the null hypothesis and p-value logic resembles a proof by contradiction (i.e., low *p*-values indicate extreme data and usually lead researchers to reject the null hypothesis and interpret this as evidence in favour of the alternative hypothesis). Unlike null-hypothesis testing, Bayesian statistics allow one to test of the degree of support for a hypothesis (i.e., degree of strength of evidence in favour of or against any given hypothesis). This is expressed as the Bayes factor (BF), which is the ratio

between the evidence in support of the alternative hypothesis over the null hypothesis (BF_{10}). By comparing the fit of the data under the null hypothesis to the alternative hypothesis, Bayes factors quantify the evidence in favour of these hypotheses (see Andraszewicz et al., 2015). For example, a Bayes factor of 10 (BF_{10} = 10) suggests that the alternative hypothesis is 10 times more likely than the null. Although Bayes factors provide a continuous measure of degree of evidence, there are some conventional approximate guidelines for interpretation (Andraszewicz et al., 2015, for a classification scheme): BF_{10} = 1 provides no evidence either way, BF_{10} > 1 anecdotal, BF_{10} > 3 moderate, BF_{10} > 10 strong, BF_{10} > 30 very strong and BF_{10} >100 decisive evidence for the alternative hypothesis. By adding these Bayesian analyses, we deepened our findings from the traditional analyses, as we were able to identify which hypothesis is most plausible (i.e., alternative hypothesis vs. null hypothesis) and which predictors are the strongest.

Results

Descriptive statistics

The descriptive statistics of the primary measures are presented in Table 1. A table with descriptive statistics of all administered measures – including control variables and baseline conditions – can be found in Appendix A.

Table 1

Descriptive statistics of the key variables

	n	*M*	*SD*	Range	Reliability
Arithmetic					
Addition					
Accuracy	127	.94	.07	[.66-1]	.95
Response time (ms)	127	3654.22	1445.67	[1716.01-8948.59]	.80

Multiplication					
Accuracy	127	.84	.12	[.36-1]	.94
Response time (ms)	127	7141.84	4134.53	[2484.21-31130.19]	.87
Executive functions					
Inhibition					
Flanker task – incongruent condition					
Accuracy	126	.91	.13	[.15-1]	.79
Response time (ms)	126	956.12	236.39	[560.60-1729.25]	.90
Inverse efficiency (RT/accuracy) [a]	126	1115.01	550.80	[604.50-5047.67]	.
Animal Stroop task – incongruent condition					
Accuracy	126	.91	.09	[.60-1]	.58
Response time (ms)	126	1000.10	167.46	[619.25-1393.40]	.80
Inverse efficiency (RT/accuracy)[a]	126	1106.09	196.96	[709.26-1732.23]	.
Shifting					
WCST (average # items needed to switch) [b]	123	7.03	5.01	[2-33]	.
Updating					
2-back task (accuracy)	127	.72	.09	[.15-.85]	.62
Metacognition					

General metacognitive knowledge					
General metacognitive questionnaire (# correct) [c]	127	9.20	2.47	[3-15]	.49
Calibration of confidence [d]					
In addition task	127	1.85	.15	[1.31-2]	.78
In multiplication task	127	1.74	.18	[0.81-2]	.74
Numerical magnitude processing					
Symbolic numerical magnitude comparison task					
Response time (ms)	127	856.51	223.15	[465.73-2037.15]	.92

Note. [a] Inverse efficiency scores were calculated by dividing the response time by the accuracy; the higher the score, the worse the performance; [b] Score = total number of items needed to switch divided by number of blocks completed; [c] Number of correct answers; [d] Alignment between children's confidence rating and the accuracy of their arithmetic answer, i.e., correct arithmetic answers yielded a score of 2 if children said they were *Correct*, 1 if they said *I am not sure*, and 0 if they said they were *Incorrect*; this scale was reversed when the arithmetic answer was incorrect. The higher the score, the better the calibration of confidence.

Preliminary analyses

Before conducting the subsequent correlation and regression analyses, we conducted a series of preliminary analyses (e.g., strategy use; performance measures in the inhibition tasks) to rule out potential alternative explanations for the current results. These analyses are included in the Supplementary material section. None of these alternative explanations accounted for the findings that are reported below.

Correlational analyses

Pearson correlation coefficients were calculated to examine the associations between the different variables under study and addition (i.e., response time (RT) and accuracy; Table 2a), and multiplication performance (i.e., RT and accuracy; Table 2b). A full matrix of all intercorrelations is provided in Appendix B. For the Bayesian analyses, we used a default prior provided by the statistical program JASP (JASP Team, 2017). The default prior width was set to 1 for Pearson correlations.

Table 2a

Correlational analyses of the response time (RT) and accuracy of addition

Variable	Addition RT			Addition accuracy		
	r	*p*	BF_{10}	*r*	*p*	BF_{10}
Executive functions						
Inhibition – Flanker (IE) [a,b]	-.156	.083	0.501	-.040	.659	0.123
Inhibition – Stroop (IE) [a, c]	.120	.183	0.270	-.188	.035	1.002
Shifting [a] (average # items needed to switch)	-.050	.585	0.131	.002	.987	0.113
Updating (accuracy) [d]	-.101	.259	0.208	.341	<.001	>100
Metacognition						
General metacognitive konwledge (# correct) [d]	-.243	.006	4.701	.114	.201	0.249
Calibration of Confidence – Addition [d]	-.289	.001	23.385	.502	<.001	>100
Symbolic numerical magnitude processing (RT) [a, e]	.361	<.001	>100	.028	.757	0.116

Note. [a] The higher the scores, the worse the performance; [b] Controlled for the Flanker task baseline condition (IE); [c] Controlled for the animal Stroop task baseline condition (IE); [d] The higher the score, the better the performance; [e] Controlled for performance on the motor speed task (RT).

Table 2b

Correlational analyses of the response time (RT) and accuracy of multiplication

Variable	Multiplication RT			Multiplication accuracy		
	r	*p*	BF_{10}	*r*	*p*	BF_{10}
Executive functions						
Inhibition – Flanker (IE) [a,b]	-.142	.114	0.386	.085	.349	0.173
Inhibition – Stroop (IE) [a, c]	.001	.981	0.111	-.081	.371	0.167
Shifting [a] (average # items needed to switch)	-.090	.324	0.182	-.060	.508	0.140
Updating (accuracy) [d]	.053	.556	0.132	.278	.002	15.478
Metacognition						
General metacognitive konwledge (# correct) [d]	-.147	.099	0.426	.077	.391	0.160
Calibration of Confidence – Addition [d]	-.238	.007	4.058	.791	<.001	>100
Symbolic numerical magnitude processing (RT) [a, e]	.251	.005	6.086	.008	.929	0.111

Note. [a] The higher the scores, the worse the performance; [b] Controlled for the Flanker task baseline condition (IE); [c] Controlled for the animal Stroop task baseline condition (IE); [d] The higher the score, the better the performance; [e] Controlled for performance on the motor speed task (RT).

Executive functions.

Performance (IE) on the Animal Stroop task was significantly correlated with addition accuracy, indicating that children with better inhibition skills performed more accurate on the addition task. The Bayes factor indicated only anecdotal evidence for this association. The n-back task (accuracy)

correlated with both addition and multiplication accuracy with Bayes factors indicating strong to decisive evidence. This correlation was driven by omission errors (and not commission errors) in the n-back task. There were no other significant correlations with the remaining executive functioning variables. Moreover, the Bayes factors indicated evidence in favour of the null hypotheses for these remaining associations.

Metacognition.

Calibration of confidence was significantly correlated with addition and multiplication performance, indicating that children with better on-task metacognitive monitoring skills performed better (i.e., faster and more accurate) on the arithmetic tasks. Bayes factors indicated moderate to decisive evidence for these associations. General metacognitive knowledge was significantly associated with addition RT, indicating that children with better global metacognitive knowledge performed faster on the addition task. The Bayes factor indicated moderate evidence in favour of this association. There was no evidence for the null nor for the alternative hypothesis for the association between multiplication RT and general metacognitive knowledge. The null hypotheses of no association were supported for the associations between general metacognitive knowledge and addition and multiplication accuracy.

***Symbolic numerical magnitude processing*.**

Addition and multiplication RT were significantly related to symbolic numerical magnitude processing: Children with better symbolic numerical magnitude processing skills performed faster when doing arithmetic. Bayes factors indicated moderate to decisive evidence. The null hypotheses were supported for the associations between addition and multiplication accuracy and symbolic numerical magnitude processing RT.

Regression analyses

Regression analyses were performed to assess the unique contribution of our different cognitive variables to arithmetic performance. Therefore, all variables that were significantly related to addition and/or multiplication were entered simultaneously into the regression model. To quantify the evidence in favour of our hypotheses, we calculated Bayes factors for each predictor and identified the best model to predict each dependent variable in JASP (JASP, 2017). Specifically, a $BF_{inclusion}$ was calculated for every predictor in the model, which represents the change from prior to posterior odds (i.e., BF_{10}), where the odds concern all the models with a predictor of interest to all models without that predictor (i.e., a Bayes factor for including a predictor averaged across the models under consideration). We used a default prior width provided by JASP of .354 (prior for *r* scale covariates) for the linear regression analyses. Table 3 presents the results of our regression analyses.

Table 3a

Regression analysis of addition RT (R^2 = .225)

Variable	*β*	*t*	*p*	ΔR^2	$BF_{inclusion}$
Control variables					
Motor speed task (RT) [a]	-.114	-1.321	.189	.024	0.474
Intelligence (# correct) [b]	.020	0.247	.806	0	0.254
Primary variables					
General metacognitive knowledge (# correct) [b]	-.178	-2.143	.034	.028	1.493
Calibration of Confidence – Addition [b]	-.291	-3.651	<.001	.082	76.675
Symbolic numerical magnitude processing (RT) [a]	.356	4.080	<.001	.103	>100

Note. [a] The higher the score, the worse the performance; [b] The higher the score, the better the performance.

Table 3b

Regression analysis of addition accuracy (R^2 = .338)

Variable	*β*	*t*	*p*	ΔR^2	$BF_{inclusion}$
Control variables					
Stroop task Baseline (IE) [a]	.023	0.250	.803	0	0.224
Intelligence (# correct) [b]	.176	2.237	.027	.028	3.885
Primary variables					
Inhibition – Stroop task (IE) [a]	-.071	-0.731	.466	.003	0.300
Updating (accuracy) [b]	.202	2.576	.011	.037	4.673
Calibration of Confidence – Addition [b]	.414	5.238	<.001	.151	>100

Note. [a] The higher the scores, the worse the performance; [b] The higher the score, the better the performance.

Table 3c

Regression analysis of multiplication RT (R^2 = .186)

Variable	*β*	*t*	*p*	ΔR^2	$BF_{inclusion}$
Control variables					
Motor speed task (RT) [a]	-.079	-0.860	.392	.005	0.36
Intelligence (# correct) [b]	.223	2.635	.010	.047	4.84
Primary variables					
General metacognitive knowledge (# correct) [b]	-.118	-1.358	.177	.012	0.63

Calibration of Confidence – Multiplication [b]	-.271	-3.202	.002	.069	16.74
Symbolic numerical magnitude processing (RT) [a]	.232	2.540	.012	.044	6.69

Note. [a] The higher the score, the worse the performance; [b] The higher the score, the better the performance.

Table 3d

Regression analysis of multiplication accuracy (R^2 = .661)

Variable	*β*	*t*	*p*	ΔR^2	$BF_{inclusion}$
Control variables					
Stroop task Baseline (IE) [a]	.006	0.086	.932	0	0.279
Intelligence (# correct) [b]	.119	2.115	.036	.013	1.539
Primary variables					
Inhibition – Stroop task (IE) [a]	.024	0.353	.725	.001	0.260
Updating (accuracy) [b]	.131	2.376	.019	.016	1.572
Calibration of Confidence – Multiplication [b]	.759	13.906	<.001	.547	>100

Note. [a] The higher the scores, the worse the performance; [b] The higher the score, the better the performance.

Addition.

Addition RT was significantly predicted by general metacognitive knowledge, calibration of confidence and symbolic numerical magnitude processing (RT). Bayes factors indicated very strong to decisive evidence for the latter two variables while the Bayes factor for general metacognitive knowledge indicated only anecdotal support for the alternative hypothesis. The strongest model to predict addition RT was the model including general metacognitive knowledge, calibration of

confidence and symbolic numerical magnitude processing (BF_{10} = 85387.75). The R^2 of the overall model was medium.

Addition accuracy was significantly predicted by the 2-back task (accuracy), calibration of confidence and intelligence, and the evidence for the contribution of these predictors was moderate to decisive. There was moderate evidence for the null hypothesis concerning the influence of the animal Stroop task (IE). The strongest model to predict addition accuracy was the model including intelligence, updating, and calibration of confidence (BF_{10} = 1.270e+8). The R^2 of the overall model was large.

Multiplication.

Multiplication RT was significantly predicted by calibration of confidence, symbolic numerical magnitude processing (RT) and intelligence. Bayes factors indicated moderate to strong evidence for these predictors. General metacognitive knowledge did not predict multiplication RT. The strongest model to predict multiplication RT was the model including intelligence, calibration of confidence and symbolic numerical magnitude processing (BF_{10} = 464.81). The R^2 of the overall model was medium.

Multiplication accuracy was significantly predicted by the 2-back task (accuracy), calibration of confidence and intelligence. Only for calibration of confidence, the Bayes factors indicated decisive evidence. The evidence for the contribution of the 2-back task accuracy and intelligence was only anecdotal. The null hypothesis was supported for no influence of the animal Stroop task (IE) in multiplication accuracy. The strongest model to predict multiplication accuracy was the model including updating and calibration of confidence (BF_{10} = 2.102e+25). The R^2 of the overall model was large.

Discussion

Why are some children so good in math and do others struggle their entire life with basic math competencies? A large body of literature has already identified different cognitive factors that play an

important role in mathematical development (e.g., Berch et al., 2016; Dowker, 2005; Duncan et al., 2007). These factors can be subdivided into two groups, namely domain-general cognitive factors and domain-specific cognitive factors (e.g., Geary & Moore, 2016). Existing research has largely focused on one domain-specific cognitive factor, that is numerical magnitude processing (Schneider et al., 2017), neglecting other critical domain-general cognitive factors that might play a role in (a)typical mathematical development. The current study included two domain-general factors, i.e. executive functions and metacognition, to identify their unique contribution in arithmetic in addition to each other as well as in addition to symbolic numerical magnitude processing as an important domain-specific factor. This study closed an important gap in the literature by focusing on executive functioning in a comprehensive way (i.e., including inhibition, shifting and updating) while additionally investigating the role of metacognition. By combining these two factors, this study sheds a new and important light on how these different variables co-operate in arithmetic. Additionally, to thoroughly investigate the role of executive functioning, the current study took into account different degrees of experience with arithmetic, i.e, high (addition) vs. low (multiplication), as the level of experience might moderate the association between the cognitive factors with arithmetic (e.g., Laski & Dulaney, 2015). On the other hand, we included both general metacognitive knowledge and on-task metacognitive monitoring to investigate to which extent domain-general as well as domain-specific measures of metacognition were related to arithmetic.

Metacognition, and specifically on-task confidence judgements, was found to be a strong, stable and unique predictor of both addition and multiplication. In line with Rinne and Mazzocco (2014), we found that local confidence judgements were strongly related with concurrent mental arithmetic. By uncovering its unique role in arithmetic above executive functions and symbolic numerical magnitude processing, and by testing this association in early primary school children, we extended Rinne and Mazzocco's findings in an important way. The current cross-sectional data do not allow us to derive strong conclusions about the direction of the association between metacognition and arithmetic. It might be that better performance on arithmetic has a positive influence on

metacognitive skills, namely that more experience with correct arithmetic answers (i.e., due to better arithmetic skills) makes detecting incorrect answers easier, which leads to better later metacognitive performance. This hypothesis is in line with the results of Roebers and Spiess (2017), who found that earlier domain-specific skills in the first-order task (i.e., spelling) predicted later metacognitive monitoring abilities. On the other hand, it might be that better detection of errors (i.e., due to better metacognitive skills) creates more learning moments in which incorrect arithmetic answers are detected and new, potentially correct arithmetic answers, are learned, which in turn leads to better later arithmetic performance. This hypothesis is in line with the results of Rinne and Mazzocco (2014), who found that better calibration predicted later gains in mental arithmetic. Lastly, the relationship between metacognition and arithmetic performance might also be reciprocal, in which arithmetic performance and metacognition both influence each other. Future longitudinal studies are needed to investigate this.

In contrast to the associations between on-task metacognitive skills and arithmetic that were unique and strong, general metacognitive knowledge was significantly correlated with addition RT, but once symbolic numerical magnitude processing RT was considered, there was no clear evidence for an effect of this factor. This might suggest that children's metacognition is more domain-specific, which is in line with observations by Vo and colleagues (2014) and Neuenhaus, Artelt, Lingel and Schneider (2011). Importantly, the general metacognitive questionnaire differed in two considerable ways from the local, on-task metacognitive monitoring measure. Compared to the general metacognitive questionnaire, on-task metacognitive monitoring was measured online and in more detail (i.e., on a trial-by-trial basis). This could limit the interpretation of specificity of metacognition based on our results. The obtained results might be specific to the age of our participants, as it is theoretically assumed that the development of metacognitive knowledge and skills begin highly domain and situation-specific and become more flexible and domain-general with practice and experience (Borkowski, Chan, & Muthukrishna, 2000).

This study also investigated different components of executive functioning, namely inhibition, shifting and updating. Based on the lack of significant correlations between our executive functioning measures (see Appendix B), our results confirm that inhibition, shifting and updating are separate aspects of executive functioning. This result is in line with existing research (e.g., Bull & Scerif, 2001; Cragg & Gilmore, 2014; Lee et al., 2013; Miyake et al., 2000; van der Sluis et al., 2007), and indicates that in the operationalisation of executive functions, it is essential to differentiate between different components.

In line with previous research (e.g., Bull & Scerif, 2001; Cragg & Gilmore, 2014; Peng et al., 2016), we found that updating accuracy was a stable and unique predictor of both addition and multiplication accuracy. Updating skills may be required to recall arithmetic facts from long-term memory, in line with evidence suggesting that individuals with low updating skills are less likely to retrieve answers to simple arithmetic problems (e.g., Barrouillet & Lépine, 2005; Geary, Hoard, & Bailey, 2012). On the other hand, arithmetic accuracy may rely on updating skills in order to represent the arithmetic item and to store interim solutions while using procedural strategies in multi-step problems.

Whereas our frequentist analyses pointed to a significant – yet weak – association between addition accuracy and inhibition, the Bayes factor indicated that the hypothesis of an association can neither be confirmed nor rejected. Consequently, based on on the results of the current study, no firm conclusions can be drawn on the association between inhibition and addition accuracy. On the other hand, the Bayes Factors of the association between inhibition and multiplication indicate there is evidence in favour of no association. Based on the current results, there is not much support for a time-limited role of inhibition in arithmetic performance, as we originally predicted. Future research is needed to further investigate the association between inhibition and arithmetic.

We did not observe an association between shifting and arithmetic. Moreover Bayes factors indicated moderate evidence in favour of no association between these variables. This result is in line

with previous research (e.g., Bull & Lee, 2014; van der Sluis et al., 2007; Van der Ven et al., 2012; Yeniad et al., 2013), but differs from other existing literature (e.g., Bull, Johnston, & Roy, 1999). This difference could be due to the difference in the operationalization of mathematics, as different mathematical abilities could all be differently related to certain cognitive factors. On the other hand, the need for shifting (e.g., between strategies) was limited in our single-digit arithmetic task, because of the curricular focus on retrieval of single-digit arithmetic (De Smedt, 2016), and the homogeneity of the arithmetic tasks (i.e., addition and multiplication items were presented in separate tasks, requiring no shifts between operations in a task). The impact of shifting skills on arithmetic might become more clear when different math components are considered together. The absence of a significant correlation of shifting skills with arithmetic in this study does not indicate its lack of influence in arithmetic, yet it suggests that individual differences in shifting skills are not a strong explanation for individual differences in single-digit arithmetic.

While both executive functions and on-task metacognitive monitoring contributed to arithmetic performance, on-task metacognitive monitoring proved to be more important. This could be due to the fact that the sensitivity of the executive functions tasks to capture individual differences was lower, because these tasks have low between-subject variability and hence their usefulness in studying individual differences might be limited (Hedge, Powell, & Sumner, 2017).

In line with a large body of research (Schneider et al., 2017, for a meta-analysis), the current findings support the unique role of symbolic numerical magnitude processing in arithmetic. The current data clearly indicate that such associations do not merely arise as a result of a common reliance on executive functions (e.g., Gilmore et al., 2013).

The results of this study revealed that there is overlap in the cognitive processes that are associated with addition and multiplication performance (e.g., updating skills accuracy, on-task metacognitive monitoring, symbolic numerical magnitude processing RT), but that the strength of the association differed across the two operations. This indicates that subtle differences (e.g., level of

experience) within one mathematical skill (i.e., arithmetic) could account for different associations with other cognitive variables. It is important to note that the difference between addition and multiplication is not purely one of experience, and that there could also be differences in task demands. On the other hand, the kind of instruction the children received in the Belgian school system, namely the curriculum's strong emphasis on automatization of arithmetic facts for both operations ("Onderwijsdoelen Vlaanderen", 2018), makes these task demands rather similar.

Future research should investigate these associations investigated in the current study longitudinally to examine the directions of the associations and how such associations evolve over time. Supplementary reaction time based analyses (see Supplementary material 2.2) indicated that different strategies were used to solve the arithmetic items, and that indices reflective of retrieval and procedure use were both correlated with on-task metacognitive judgements and symbolic magnitude processing. However, these metrics might mask important individual differences in strategy use and future research should investigate strategy use on a trial-by-trial level, for example, by collecting verbal report data.

To conclude, the results of this study show that updating, metacognition and numerical magnitude processing are all significant predictors of arithmetic performance, even in addition to each other. Importantly, these results emphasise the importance of including metacognition in cognitive research. Metacognition has received a lot of attention in more general educational research, yet is mostly ignored in developmental research in the field of mathematics. Importantly, this study reveals that noticing your own errors (i.e., task-specific metacognitive monitoring) is an important unique predictor of arithmetic performance. Knowing about your own tendency to commit easy errors may lead to increased self-regulatory activities (Schneider, 2010), and might lead to improvements in (arithmetic) performances over time (Rinne & Mazzocco, 2014). Hence, these individual differences in noticing your own errors might help you to learn from your mistakes.

REFERENCES

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*(2), 521–543. https://doi.org/10.1177/0149206314560412

Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology*, *8*(4), 485–493.

Barrouillet, P., & Lépine, R. (2005). Working memory and children' s use of retrieval to solve addition problems. *Journal of Experimental Child Psychology*, *91*, 183–204. https://doi.org/10.1016/j.jecp.2005.03.002

Berch, D. B., Geary, D. C., & Mann Koepke, K. (2016). *Development of mathematical cognition: neural substrates and genetic influences.*

Borkowski, J. G., Chan, L. K. S., & Muthukrishna, N. (2000). A process-oriented model of metacognition: Links between motivation and executive functioning of metacognition. *Issues in the Measurement of Metacognition*, 1–41. Retrieved from http://digitalcommons.unl.edu/burosmetacognition/2/

Bryce, D., Whitebread, D., & Szűcs, D. (2015). The relationships among executive functions, metacognitive skills and educational achievement in 5 and 7 year-old children. *Metacognition Learning*, *10*, 181–198. https://doi.org/10.1007/s11409-014-9120-4

Bull, R., Johnston, R. S., & Roy, J. a. (1999). Exploring the roles of the visual-spatial sketch pad and central executive in children's arithmetical skills: Views from cognition and developmental neuropsychology. *Developmental Neuropsychology*, *15*(3), 421–442. https://doi.org/10.1080/87565649909540759

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*(1), 36–41. https://doi.org/10.1111/cdep.12059

Bull, R., & Scerif, G. (2001). Executive functionin as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*(3), 273–293. https://doi.org/10.1207/S15326942DN1903

Carlson, S. M., Zelazo, P. D., & Faja, S. (2013). Executive Function. In P. D. Zelazo (Ed.), *The Oxford Handbook of Developmental Psychology* (pp. 706–743). https://doi.org/10.1093/oxfordhb/9780199958450.013.0025

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63–68. https://doi.org/10.1016/j.tine.2013.12.001

Cragg, L., Keeble, S., Richardson, S., Roome, H. E., & Gilmore, C. (2017). Direct and indirect influences of executive functions on mathematics achievement. *Cognition*, *162*, 12–26. https://doi.org/10.1016/j.cognition.2017.01.014

De Smedt, B. (2016). Individual differences in arithmetic fact retrieval. In D. C. Geary & K. Mann-Koepke (Eds.), *Development of mathematical cognition: Neural substrates and genetic influences* (pp. 219–243). San Diego, CA: Elsevier Academic Press.

De Smedt, B., & Boets, B. (2010). Phonological processing and arithmetic fact retrieval: Evidence from developmental dyslexia. *Neuropsychologia*, *48*(14), 3973–3981. https://doi.org/10.1016/j.neuropsychologia.2010.10.018

De Smedt, B., Noël, M.-P., Gilmore, C., & Ansari, D. (2013). The relationship between symbolic and non-symbolic numerical magnitude processing skills and the typical and atypical development of mathematics: A review of evidence from brain and behavior. *Trends in Neuroscience & Education*, *2*, 48–55. https://doi.org/10.1023/A

De Visscher, A., & Noël, M.-P. (2014a). The detrimental effect of interference in multiplication facts storing: Typical development and individual differences. *Journal of Experimental Psychology:*

General, *143*(6), 2380–2400. https://doi.org/10.1037/xge0000029

De Visscher, A., & Noël, M. P. (2014b). Arithmetic facts storage deficit: The hypersensitivity-to-interference in memory hypothesis. *Developmental Science*, *17*(3), 434–442. https://doi.org/10.1111/desc.12135

Desender, K., Van Opstal, F., & Van den Bussche, E. (2014). Feeling the conflict: The crucial role of conflict experience in adaptation. *Psychological Science*, *25*(3), 675–683. https://doi.org/10.1177/0956797613511468

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dowker, A. (2005). *Individual differences in arithmetic. Implications for psychology, neuroscience and education.* Hove: Psychology Press.

Duncan, G. J., Claessens, A., Magnuson, K., Klebanov, P., Pagani, L. S., Feinstein, L., … Japel, C. (2007). School Readiness and Later Achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Fias, W., Menon, V., & Szucs, D. (2013). Multiple components of developmental dyscalculia. *Trends in Neuroscience and Education*, *2*(2), 43–47. https://doi.org/10.1016/j.tine.2013.06.006

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring A New Area of Cognitive — Developmental Inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066x.34.10.906

Fleming, S. M., & Dolan, R. J. (2014). The Neural Basis of Metacognitive Ability. In S. M. Fleming & C.

D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 245–265). Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-45190

Freeman, E. E., Karayanidis, F., & Chalmers, K. A. (2017). Metacognitive monitoring of working memory performance and its relationship to academic achievement in Grade 4 children. *Learning and Individual Differences*, *57*, 58–64. https://doi.org/10.1016/j.lindif.2017.06.003

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*. https://doi.org/10.1016/j.cortex.2016.04.023

Friso-Van Den Bos, I., Van Der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44. https://doi.org/10.1016/j.edurev.2013.05.003

Geary, D. C., Hoard, M. K., & Bailey, D. H. (2012). Fact retrieval deficits in low achieving children and children with mathematical learning disability. *Journal of Learning Disabilities*, *45*(4), 291–307. https://doi.org/10.1177/0022219410392046

Geary, D. C., & Moore, A. M. (2016). Cognitive and brain systems underlying early mathematical development. *Progress in Brain Research*, *227*, 75–103. https://doi.org/10.1016/bs.pbr.2016.03.008

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., … Inglis, M. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. *PLoS ONE*, *8*(6), 1–9. https://doi.org/10.1371/journal.pone.0067374

Gilmore, C., Keeble, S., Richardson, S., & Cragg, L. (2015). The role of cognitive inhibition in different components of arithmetic. *ZDM Mathematics Education*, 1–12. https://doi.org/10.1007/s11858-014-0659-y

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting

to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404–411. https://doi.org/10.1037/h0059831

Haberkorn, K., Lockl, K., Pohl, S., Ebert, S., & Weinert, S. (2014). Metacognitive knowledge in children at early elementary school. *Metacognition and Learning*, *9*(3), 239–263. https://doi.org/10.1007/s11409-014-9115-1

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 1–21. https://doi.org/10.3758/s13428-017-0935-1

Huizinga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, *44*(11), 2017–2036. https://doi.org/10.1016/j.neuropsychologia.2006.01.010

JASP (2017). JASP. Retrieved from https://jasp-stats.org/

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding It Up: Helping Children Learn Mathematics*. Washington, DC: National Academy Press. https://doi.org/10.17226/9822

Laski, E. V, & Dulaney, A. (2015). When prior knowledge interferes, inhibitory control matters for learning: The case of numerical magnitude representations. *Journal of Educational Psychology*.

Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development*, *84*(6). https://doi.org/10.1111/cdev.12096

Lyons, K. E., & Zelazo, P. D. (2011). Monitoring, metacognition, and executive function: Elucidating the role of self-reflection in the development of self-regulation. In B. B. Janette (Ed.), *Advances in child development and behovior* (pp. 379–412). Oxford, United Kingdom: Elsevier.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex Frontal Lobe

Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*, 49–100. https://doi.org/10.1006/cogp.1999.0734

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326.

Nelson, T. 0, & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125–173. Retrieved from http://www.imbs.uci.edu/~lnarens/1990/Nelson&Narens_Book_Chapter_1990.pdf

Neuenhaus, N., Artelt, C., Lingel, K., & Schneider, W. (2011). Fifth graders metacognitive knowledge: General or domain-specific? *European Journal of Psychology of Education*, *26*(2), 163–178. https://doi.org/10.1007/s10212-010-0040-7

Onderwijsdoelen Vlaanderen. (2018). Retrieved from https://www.onderwijsdoelen.be/wiskunde-lager-onderwijs

Pelegrina, S., Lechuga, M. T., García-Madruga, J. A., Elosúa, M. R., Macizo, P., Carreiras, M., … Bajo, M. T. (2015). Normative data on the n-back task for children and young adolescents. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.01544

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, *108*(4), 455–473. https://doi.org/10.1037/edu0000079

Pieschl, S. (2009). Metacognitive calibration-an extended conceptualization and potential applications. *Metacognition and Learning*, *4*(1), 3–31. https://doi.org/10.1007/s11409-008-9030-4

Raven, C. J., Court, J. H., & Raven, J. (1992). *Standard progressive matrices.* Oxford, United Kingdom: Oxford Psychologist Press.

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS ONE*, *9*(7), 1–11. https://doi.org/10.1371/journal.pone.0098663

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, *45*, 31–51. Retrieved from http://ac.els-cdn.com/S0273229716300132/1-s2.0-S0273229716300132-main.pdf?_tid=0a1eaf4e-6545-11e7-9b40-00000aacb35d&acdnat=1499673465_6d5a90fcf679237c7c78357ddf9a05ff

Roebers, C. M., & Feurer, E. (2015). Linking Executive Functions and Procedural Metacognition. *Child Development Perspectives*, *10*(1), 39–44. https://doi.org/10.1111/cdep.12159

Roebers, C., & Spiess, M. (2017). The Development of Metacognitive Monitoring and Control in Second Graders: A Short-Term Longitudinal Study. *Journal of Cognition and Development*, *18*(1), 110–128. https://doi.org/10.1080/15248372.2016.1157079

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*, *20*(3), 1–16. https://doi.org/10.1111/desc.12372

Schneider, W. (2010). The development of metacognitive competences. In B. M. Glatzeder, A. von Müller, & V. Goel (Eds.), *Towards a theory of thinking* (pp. 203–214). Heidelberg: Springer-Verlag. Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-3-642-03129-8_14.pdf

Schneider, W. (2015). *Memory Development from Early Childhood Through Emerging Adulthood*. Cham: Springer. Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-3-319-09611-7.pdf

Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *ZDM - International*

Journal on Mathematics Education, *42*, 149–161. https://doi.org/10.1007/s11858-010-0240-2

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: evidence for developmental trends. In J. Dunlosky & R. A. Bjork (Eds.), *A handbook of metamemory and memory*. Mahwah, NJ: Erlbaum.

Shrager, J., & Siegler, R. S. (1998). A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*(5), 405–410.

Siegler, R. S. (1996). *Emerging Minds. The process of change in children's thinking*. Oxford University Press.

Szucs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2013). Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. *Cortex*, *49*(10), 2674–2688. https://doi.org/10.1016/j.cortex.2013.06.007

van der Sluis, S., de Jong, P. F., & van der Leij, A. (2007). Executive functioning in children, and its relations with reasoning, reading, and arithmetic. *Intelligence*, *35*(5), 427–449. https://doi.org/10.1016/j.intell.2006.09.001

Van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012). The development of executive functions and early mathematics : A dynamic relationship. *British Journal of Educational Psychology*, *82*, 100–119. https://doi.org/10.1111/j.2044-8279.2011.02035.x

Vanbinst, K., Ceulemans, E., Ghesquière, P., & De Smedt, B. (2015). Profiles of children's arithmetic fact development: A model-based clustering approach. *Journal of Experimental Child Psychology*, *133*, 29–46. https://doi.org/10.1016/j.jecp.2015.01.003

Vanbinst, K., & De Smedt, B. (2016). Individual differences in children's mathematics achievement: The roles of symbolic numerical magnitude processing and domain-general cognitive functions. *Progress in Brain Research*, *227*, 105–130. https://doi.org/10.1016/bs.pbr.2016.04.001

Verguts, T., & Fias, W. (2005). Interacting neighbors: A connectionist model of retrieval in single-digit multiplication. *Memory & Cognition*, *33*(1), 1–16. https://doi.org/10.3758/BF03195293

Vo, V. A., Li, R., Kornell, N., Pouget, A., & Cantlon, J. F. (2014). Young children bet on their numerical skills: Metacognition in the numerical domain. *Psychological Science*, *25*(9), 1712–1721. https://doi.org/10.1177/0956797614538458

Watson, T. D., Azizian, A., & Squires, N. K. (2006). Event-related potential correlates of extradimensional and intradimensional set-shifts in a modified Wisconsin Card Sorting Test. *Brain Research*, *1092*(1), 138–151. https://doi.org/10.1016/j.brainres.2006.03.098

Yeniad, N., Malda, M., Mesman, J., Van Ijzendoorn, M. H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences*, *23*, 1–9. https://doi.org/10.1016/j.lindif.2012.10.004

Appendix

Appendix A

Descriptive statistics of all the administered measures.

Table A1

Descriptive statistics of all the administered measures

	n	*M*	*SD*	Range
Arithmetic				
Addition				
Accuracy	127	.94	.07	[.66-1]
Response time (ms)	127	3654.22	1445.67	[1716.01-8948.59]
Multiplication				
Accuracy	127	.84	.12	[.36-1]
Response time (ms)	127	7141.84	4134.53	[2484.21-31130.19]
Executive functions				
Inhibition				
Flanker – incongruent condition				
Accuracy	126	.13	.91	[.15-1]
Response time (ms)	126	956.12	236.39	[560.60-1729.25]
Inverse efficiency [a]	126	1115.01	550.80	[604.50-5047.67]
Flanker – baseline condition				
Accuracy	126	.94	.08	[.60-1]
Response time (ms)	126	647.82	87.39	[453.00-887.78]

Inverse efficiency [a]	126	689.36	101.26	[479.80-986.42]
Animal Stroop – incongruent condition				
Accuracy	126	.91	.09	[.60-1]
Response time (ms)	126	1000.10	167.46	[619.25-1393.40]
Inverse efficiency [a]	126	1106.09	196.96	[709.26-1732.23]
Animal Stroop – baseline condition				
Accuracy	126	.97	.05	[.70-1]
Response time (ms)	126	867.47	155.13	[508.05-1297.38]
Inverse efficiency [a]	126	895.11	182.83	[508.05-1651.02]
Shifting				
WCST (average # items needed to switch) [b]	123	7.03	5.01	[2-33]
Updating				
2-back task (accuracy)	127	.72	.09	[.15-.85]
Metacognition				
General metacognitive knowledge				
General metacognitive questionnaire (# correct) [c]	127	9.20	2.47	[3-15]
Calibration of confidence [d]				
In addition task	127	1.85	.15	[1.31-2]
In multiplication task	127	1.74	.18	[0.81-2]
Numerical magnitude processing				
Symbolic numerical magnitude comparison task				
Response time (ms)	127	856.51	223.15	[465.73-2037.15]
Control variables				
Intelligence (# correct) [c]	127	34.5	8.35	[10-50]
Motor speed task (RT)	126	593.86	153.92	[339.75-1151.56]

Note. [a] Calculated by dividing the RT by the accuracy; the higher the scores, the worse the performance; [b] Score = total number of items needed to switch divided by number of blocks completed; [c] Number of correct answers; [d] Alignment between children's confidence rating and the accuracy of their arithmetic answer, i.e., correct arithmetic answers yielded a score of 2 if children said they were *Correct*, 1 if they said *I am not sure*, and 0 if they said they were *Incorrect*; this scale was reversed when the arithmetic answer was incorrect. The higher the scores, the better the calibration of confidence.

Appendix B

Correlational analyses between the administered measures.

Table B1

Correlational analyses between the administered measures

			1a	1b	2a	2b	3	4	5	6	7	8	9	10	11
	Addition	1a. Addition RT													
		1b. Addition ACC													
		r	-.247												
		p	.005												
		BF_{10}	5.317												
Arithmetic	Multiplication	2a. Multiplication RT													
		r	.558	.153											
		p	<.001	.086											
		BF_{10}	>100	0.476											
		2b. Multiplication ACC													
		r	-.525	.413	-.168										
		p	<.001	<.001	.058										
		BF_{10}	>100	>100	0.653										
	Inhibition	3. Flanker task IE													
		r	-.156	-.040	-.142	.085									
		p	.083	.659	.114	.349									
		BF_{10}	0.501	0.123	0.386	0.173									
		4. Stroop task IE													
Executive functions		*r*	.120	-.188	.002	-.081	.220								
		p	.183	.035	.981	.371	.015								
		BF_{10}	0.270	1.002	0.111	0.167	2.254								
	Shifting	5. WCST													
		r	-.050	.002	-.090	-.060	.121	-.081							
		p	.585	.987	.324	.508	.186	.377							

		BF_{10}	0.131	0.113	0.182	0.140	0.274	0.167					
	Updating	6. 2-back task ACC											
		r	-.101	.341	.053	.278	-.118	.014	-.091				
		p	.259	<.001	.556	.002	.192	.877	.319				
		BF_{10}	0.208	>100	0.132	15.478	0.262	0.113	0.184				
	General metacognitive knowledge	7. Metacognitive questionnaire											
		r	-.243	.114	-.147	.077	-.145	-.116	.022	.084			
		p	.006	.201	.099	.391	.107	.199	.809	.345			
		BF_{10}	4.701	0.249	0.426	0.160	0.408	0.255	0.116	0.172			
Metacognition	Task-specific metacognition	8. Calibration of confidence - Addition											
		r	-.289	.502	-.030	.396	-.104	-.205	.082	.262	.065		
		p	.001	<.001	.735	<.001	.250	.022	.369	.003	.468		
		BF_{10}	23.385	>100	0.117	>100	0.216	1.531	0.168	8.843	0.144		
		9. Calibration of confidence - Multiplication											
		r	-.451	.286	-.238	.791	.039	-.092	-.082	.170	.055	.507	
		p	<.001	.001	.007	<.001	.666	.310	.367	.056	.536	<.001	
		BF_{10}	>100	20.796	4.058	>100	0.122	0.187	0.169	0.675	0.134	>100	
Domain-specific factor	Nnumerical magnitude processing	10. Symbolic numerical magnitude processing RT											
		r	.361	.028	.251	.008	-.081	-.010	-.004	.021	-.158	-.013	.052
		p	<.001	.757	.005	.929	.369	.909	.963	.816	.079	.887	.564
		BF_{10}	>100	0.116	6.086	0.111	0.167	0.112	0.113	0.114	0.516	0.112	0.131
Control Variables	Intelligence	11. Raven											
		r	.034	.275	.221	.221	-.310	-.275	-.084	.158	.079	.118	.115	.146
		p	.705	.002	.013	.012	<.001	.002	.357	.076	.379	.186	.199	.105
		BF_{10}	0.119	14.099	2.392	2.449	53.106	13.534	0.171	0.525	0.163	0.263	0.251	0.415

Motor speed	12. Motor reaction time task RT												
	r	.087	.036	.098	-.207	-.058	.117	-.136	-.186	-.242	-.083	-.184	.375	.055
	p	.331	.690	.273	.020	.521	.195	.134	.037	.006	.358	.039	<.001	.544
	BF_{10}	0.178	0.121	0.202	1.623	0.137	0.258	0.343	0.946	4.386	0.169	0.917	>100	0.134

Note. ACC = accuracy; All correlations with the Flanker task incongruent condition are controlled for the performance (IE) on the Flanker baseline condition; All correlations with the animal Stroop task incongruent condition are controlled for the performance (IE) on the animal Stroop baseline task; All correlations with the numerical magnitude processing task are controlled for performance (RT) on the motor speed task, except for the correlations with the inhibition tasks, which are controlled for their respective baseline (IE).

List of figure legends

Figure 1. Example stimuli for the animal Stroop task. Left: congruent condition; Right: incongruent condition.

Figure 2. Example of the response form of the general metacognitive knowledge questionnaire.

Supplementary material

1. Global metacognitive questions

Two global metacognitive questions were asked in both arithmetic tasks. Before the task started, children had to report how they thought they would perform on the task (i.e., prospective question). After the task was finished, they had to report how they thought they had performed on the task (i.e., retrospective question). We used emoticons corresponding with the possible answers (e.g., ☺ corresponded with "Good") to make the task more attractive and feasible for children.

Our rationale for the scoring of this task was the same as for the local, on-task metacognitive monitoring, namely giving the best score (i.e., 2) to correct extreme judgements (i.e., "Good" or "Not good"), the lowest score (i.e., 0) to incorrect extreme judgements, and an average score (i.e., 1) when the child did not make an extreme judgement. However, to define the correctness of the judgement, an indicator of actual performance has to be taken into account. In preparatory analyses, we tried to find such an indicator (1) by comparing children's performance to the performance of all participants (i.e., their percentile score), and (2) by comparing their performance to certain standards (i.e., accuracy of 1-.71 = Good; .70-.50 = Average; .50-0 = Not good). The children, however, were not aware of these indicators, and therefore they were not able to effectively rate their performance. It could be that they were thinking about "What does my teacher thinks good performance is?" or "What do I think the experimenter thinks about my performance?", or that they compared to some internal reference point (e.g., their typical performance on similar tasks). On the other hand, the use of one single item to measure a variable is not reliable.

Against this background, we decided to remove these two questions (i.e., prediction and postdiction) from further analyses.

2. Preliminary analyses

We report the outcome of four preliminary analyses that were run to rule out potential alternative explanations for the current results that are discussed in our manuscript.

2.1. Blocks in the arithmetic tasks

The results of arithmetic were based on the data of both blocks of each task, i.e., one without and one with metacognitive monitoring question. It could be that asking a metacognitive question after solving arithmetic items, had an influence on arithmetic performance. We therefore also analysed our data based only on the arithmetic block that did not include metacognitive questions. The results were identical to the results based on both blocks together. Consequently, the results presented in the manuscript are based on the full dataset (i.e., block with and without metacognitive monitoring questions).

2.2. Strategy use within the arithmetic tasks

The RTs of the arithmetic tasks indicated that it is likely that different strategies were used to solve the arithmetic items. We used different methods to estimate strategy use (i.e., procedural strategies vs. arithmetic fact retrieval). We calculated a Tau estimate which reflects the positive skew of the distribution of response times of a participant, with increases in Tau reflecting a slowing on some trials (Penner-Wilger & Lefevre, 2006). As a result, Tau can be used as an index of procedure use. On the other hand, we used the data of the small arithmetic items (i.e., items with an answer smaller or equal to 10 for the addition items, and items with an answer smaller or equal to 25 for the multiplication items) as an index of retrieval. This problem-size approach to investigate strategy use has been widely used in the literature (De Smedt, 2016) and the assumption that retrieval strategies

are typically used more often to solve small problems has been confirmed by empirical data in children (e.g., Barrouillet, Mignon, & Thevenot, 2008; Imbo & Vandierendonck, 2008). We calculated average RTs of the small items and a Tau estimate for every participant, and correlated these two with the other key variables in our study (see Table S1 and Table S2 below). Importantly, as they are based on RT, these estimates of strategy use do not provide information about the associations between our key variables and arithmetic accuracy.

Table S1

Correlational analyses of the response time (RT) of the small items as an index of retrieval and the Tau estimate as an index of procedural use in the addition task

Variable	Small			Tau		
	r	*P*	BF_{10}	*r*	*p*	BF_{10}
Executive functions						
Inhibition: Flanker (IE) [a, b]	-.051	.571	0.131	-.092	.310	0.187
Inhibition: Stroop (IE) [a, c]	.105	.243	0.219	.189	.035	1.026
Shifting (average # items needed to switch) [d]	-.006	.945	0.115	-.074	.423	0.157
Updating (accuracy)	-.120	.179	0.271	-.111	.215	0.238
Metacognition						
General metacognitive knowledge (# correct) [e]	-.218	.014	2.132	-.202	.023	1.393
Calibration of Confidence – Addition [e]	-.228	.011	2.769	-.216	.016	1.978

Symbolic numerical magnitude processing (RT) [f]	.379	<.001	>100	.258	.005	5.854

Note. [a] IE = inverse efficiency of the incongruent condition, calculated by dividing the response time by the accuracy; the higher the score, the worse the performance [b]; Controlled for the Flanker task baseline condition (IE); [c] Controlled for the animal Stroop task baseline condition (IE); [d] The higher the score, the worse the performance; [e] The higher the score, the better the performance; [f] Controlled for performance on the motor speed task (RT).

Table S2

Correlational analyses of the response time (RT) of the small items as an index of retrieval and the Tau estimate as an index of procedural use in the multiplication task

Variable	Small			Tau		
	r	*p*	BF_{10}	*r*	*p*	BF_{10}
Executive functions						
Inhibition: Flanker (IE) [a, b]	-.078	.386	0.162	-.086	.340	0.338
Inhibition: Stroop (IE) [a, c]	.041	.648	0.123	.069	.443	0.149
Shifting (average # items needed to switch) [d]	-.014	.879	0.116	-.052	.576	0.134
Updating (accuracy)	-.056	.530	0.135	-.013	.886	0.112
Metacognition						
General metacognitive knowledge (# correct) [e]	-.140	.117	0.371	-.137	.125	0.352
Calibration of Confidence – Multiplication [e]	-.513	<.001	>100	-.434	<.001	>100

Symbolic numerical magnitude processing (RT) [f]	.288	.002	16.069	.360	<.001	>100

Note. [a] IE = inverse efficiency of the incongruent condition, calculated by dividing the response time by the accuracy; the higher the score, the worse the performance [b]; Controlled for the Flanker task baseline condition (IE); [c] Controlled for the animal Stroop task baseline condition (IE); [d] The higher the score, the worse the performance; [e] The higher the score, the better the performance; [f] Controlled for performance on the motor speed task (RT).

As can be seen in tables S1 and S2, the pattern of results is largely the same for both strategy indicators in both operations. Consequently, based on these estimates, we conclude that our results were not likely to be confounded by strategy use and, therefore, no distinction in strategy use was made in the results presented in the manuscript. However, the analyses using these estimates should be interpreted with caution, as they do not consider strategy use on a trial-by-trial basis (Siegler, 1987) and children could be using other strategies than assumed based on the assumptions made to calculate these estimates. Future studies should therefore measure the strategy that participants use to solve each problem via trial-by-trial strategy reports.

2.3. Task-specific metacognitive monitoring

The score of task-specific metacognitive monitoring included accuracy, but accuracy was very high in the arithmetic tasks. As children tend to be overconfident (Destan & Roebers, 2015) - indeed also in this study children often indicated that they thought they were right, which in most cases (i.e., combined with accurate arithmetic answers) resulted in the highest metacognitive monitoring score - our results might be biased. We therefore also analysed the metacognitive monitoring data based on the items with incorrect arithmetic answers only. The interpretation of the results did not change when only the incorrectly solved arithmetic items were considered. Therefore, all results in the paper are based on the full dataset.

2.4. Inhibition tasks

In the inhibition tasks, i.e., Flanker task and animal Stroop task, there were different conditions, namely a baseline or neutral condition (i.e., Flanker: only one arrow was presented in the centre of the screen; Stroop: animals occupied the same area on the screen), a congruent condition (i.e., Flanker: distracter arrows pointed in the same direction as the target; Stroop: larger animal in real life = larger image on the screen) and an incongruent condition (i.e., Flanker: distracter arrows pointed in the opposite direction as the target; Stroop: larger animal in real life = smaller image on the screen). Descriptive statistics for the different conditions are presented in Table S3 below. To check if the task manipulation in the inhibition tasks worked, we verified if the performance (IE) on the incongruent condition differed from the performance (IE) on the baseline condition. This was the case for both the Flanker task ($t(125) = -9.03$, $p < .001$) and the animal Stroop task ($t(125) = 13.54$, $p < .001$). There was no facilitation-effect (i.e., better performance in the congruent condition than the baseline condition) in neither the Flanker task nor the animal Stroop task. Performance was even significantly poorer in the congruent condition in both the Flanker task ($t(125) = -9.772$, $p < .001$) and the animal Stroop task ($t(125) = -3.338$, $p < .001$) compared to the baseline condition. Adding another dimension to the task (i.e., different picture sizes) clearly made the task more difficult in both the congruent and incongruent conditions. This might be due to inhibitory demands in both conditions and therefore, we did not use the difference score (i.e., incongruent condition minus congruent condition) as index for inhibition. For all analyses presented in the manuscript, we used the IE scores of the incongruent condition controlled for the IE score of their respective baseline condition via residuals using partial correlation.

Table S3

Descriptive statistics of the different conditions (i.e., neutral, congruent and incongruent) in the inhibition tasks

	N	*M*	*SD*	Range
Flanker task				
Baseline task				
Neutral condition (IE) [a,b]	216	689.36	101.26	[479.80-986.42]
Inhibitory task				
Congruent condition (IE) [a,c]	126	840.16	200.15	[493-1730.53]
Incongruent condition (IE) [a,d]	126	1115.01	550.80	[604.50-5047.67]
Animal Stroop task				
Baseline task				
Neutral condition (IE) [a,e]	126	895.11	185.83	[508.05-1651.02]
Inhibitory task				
Congruent condition (IE) [a,f]	126	942.49	208.66	[560.53-1790.44]
Incongruent condition (IE) [a,g]	126	1106.09	196.96	[709-1732.23]

Note. [a] IE = inverse efficiency, calculated by dividing the response time by the accuracy; the higher the scores, the worse the performance; [b] One arrow in the middle of the screen; [c] Distractor arrows in the same direction as the target; [d] Distractor arrows in the opposite direction as the target; [e] Both animals occupied the same area on the screen; [f] Larger animal in real life = larger image on the screen; [g] Larger animal in real life = smaller image on the screen.