Learning and Instruction 40 (2015) 1–8

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Evidence for children's error sensitivity during arithmetic word problem solving

Amélie Lubin [a,b,c,d,*], Olivier Houdé [a,b,c,d,e], Wim de Neys [a,b,c,d]

[a] CNRS, UMR 8240, LaPsyDÉ, France
[b] Université Paris Descartes, Sorbonne Paris Cité, UMR 8240, LaPsyDÉ, France
[c] Université de Caen Basse-Normandie, UMR 8240, LaPsyDÉ, France
[d] Normandie Université, UMR 8240, LaPsyDÉ, France
[e] Institut Universitaire de France, France

ARTICLE INFO

Article history:
Received 16 March 2015
Received in revised form
9 June 2015
Accepted 26 July 2015
Available online xxx

Keywords:
Arithmetic word problems
Conflict detection
Heuristic
Response confidence
School-aged children

ABSTRACT

Solving simple arithmetic word problems is often challenging for children. Recent research suggests that children often fail to solve certain of these problems because they fail to inhibit erroneous heuristic intuitions that bias their judgment. However it is unclear whether these errors result from an error monitoring or inhibition failure. Our study focuses on this critical error detection. Eight to eleven year-old schoolchildren were given problems in which an intuitively cued heuristic answer conflicted with the correct answer and a control version in which this conflict was not present. After solving each version children were asked to indicate their response confidence. Results showed that children showed a sharp confidence decrease after having failed to solve the conflict problems. This indicates that erring children have some minimal awareness of the questionable nature of their answer and underscores that they have more arithmetic understanding than their errors might seem to suggest.

1. Introduction

Solving simple arithmetic word problems is a key ability that children need to master throughout their elementary school mathematics curriculum. These simple word problems involve basic mathematical operations such as addition and subtraction. Although even young infants have been shown to have precocious knowledge of elementary arithmetic operations (Lubin, Poirel, Rossi, Pineau, & Houdé, 2009; Wynn, 1992), solving arithmetic word problems is often challenging for school-aged children and even for adults (Verschaffel, 1994). In arithmetic word problems, compare problems are typically considered to be the most difficult[1] (e.g., De Corte & Verschaffel, 1987; Giroux & Ste-Marie, 2001; Lewis & Mayer, 1987; Morales, Shute, & Pellegrino, 1985; Riley & Greeno, 1988; Schumacher & Fuchs, 2012; Stern, 1993). Consider the following example (Riley, Greeno, & Heller, 1983):

> Mary has 8 marbles. She has 5 more marbles than John. How many marbles does John have?

What makes these problems hard is that they introduce relational terminology such as “less than” or “more than” (Schumacher & Fuchs, 2012). In addition, as the introductory problem illustrates, the relational term that is introduced can be inconsistent with the arithmetic operation (e.g., subtraction) required to solve the problem (Lewis & Mayer, 1987; called these “inconsistent language” problems). Hence, the relational term will cue a response that conflicts with the correct mathematical response. That is, children will be tempted to add rather than to subtract (e.g., they will answer “13” instead of “3”). The available evidence indeed suggests that the incorrect responses in these type of problems are typically so-called “reversal errors” characterized by adding the numbers instead of subtracting them or vice versa (Lewis & Mayer, 1987; Stern, 1993; Stern & Lehrndorfer, 1992; Verschaffel, de Corte, & Pauwels, 1992). The aim of the present study is to better understand the nature of these errors in this type of arithmetic word problem in elementary schoolchildren.

Recently, Lubin, Vidal, Lanoë, Houdé, and Borst (2013) suggested that failures to solve the problems are related to an executive failure

* Corresponding author. Université Paris Descartes, Sorbonne, 46 rue Saint Jacques, 75005 Paris, France.
E-mail address: amelie.lubin@parisdescartes.fr (A. Lubin).

[1] This specifically applies to the additive word problems (e.g., De Corte & Verschaffel, 1987; Giroux & Ste-Marie, 2001; Lewis & Mayer, 1987; Morales et al., 1985; Riley & Greeno, 1988; Schumacher & Fuchs, 2012; Stern, 1993).

http://dx.doi.org/10.1016/j.learninstruc.2015.07.005

to inhibit an overlearned arithmetic strategy or heuristic. They hypothesized that errors occur because children will intuitively rely on an automatically activated “add if more, subtract if less” rule of thumb or heuristic. Interestingly, this intuitive strategy emerges precociously, is reinforced by academic learning, and is still present in adulthood (De Corte, Verschaffel, & Pauwels, 1990; Tirosh, Tsamir, & Hershkovitz, 2008; Vamvakoussi, Van Dooren, & Verschaffel, 2012). Note that the “add if more, subtract if less” heuristic can be considered as a special case of the “key-word” strategy (e.g., De Corte et al., 1990; Hegarty, Mayer, & Green, 1992; Stern, 1993; Verschaffel, 1994; Verschaffel et al., 1992). The key-word strategy refers to a general tendency whereby children base their choice of strategy (i.e., to add or subtract) on a superficial look at the key word in the problem statement (e.g., “more/less” in the examples here but more generally also related words such as “win/lose” or “gain/loss”).

Clearly, in and by itself, in many cases the “add if more, subtract if less” heuristic (or key word strategy) can be useful and will help children to arrive at a correct response. Consider the following example in which the relational term and required mathematical operation are consistent (Lewis & Mayer, 1987 called this a “consistent language problem”):

> Mary has 8 marbles. John has 5 more marbles than Mary. How many marbles does John have?

In this case applying the heuristic will cue the correct answer “13”. However, the point is that sometimes (i.e., when the relational term is inconsistent with the required mathematical operation) it will also cue a response that conflicts with the correct mathematical answer and bias our reasoning. Consequently, correctly solving such “conflict” problems will require that children inhibit the tendency to simply apply the heuristic.

To validate their claim about the role of inhibitory processing in avoiding arithmetic word problem errors, Lubin et al. (2013) adopted a negative priming paradigm (Tipper, 1985). The basic idea behind this paradigm is simple: if you inhibit a specific strategy on one trial, then activation of this same strategy on a subsequent trial should be hampered (Borst, Moutier, & Houdé, 2013). Bluntly put, when you block a strategy at Time 1, you will pay a price when trying to reactivate it again immediately afterwards. Therefore, Lubin et al. had children first solve a “conflict” arithmetic word problem in which they needed to refrain from using the “add if more, subtract if less” heuristic (e.g., the relational term and required mathematical operation were inconsistent, e.g., “Mary has 8 marbles. She has 5 more marbles than John. Does John have 13 marbles?”). Immediately afterwards they were presented with a “no-conflict” arithmetic word problem in which the heuristic cued the correct response (i.e., the relational term and required operation were consistent, e.g., “Mary has 8 marbles. John has 5 more marbles than Mary. Does John have 13 marbles?”). Lubin et al. observed that sixth-graders, nine-graders and adults were slowed down on the no-conflict problem when they had previously solved the conflict problem correctly. When the no-conflict problem was preceded by a control problem that did not require blocking the heuristic (e.g., “Joe has 25 pens. Marc has 10 pens. Does Joe have more pens than Marc?”) such slowing down was not observed. This pattern is consistent with the postulated role of inhibitory processing in arithmetic word problem solving.

In general, accounts that have stressed the importance of inhibition in human cognition and development have received wide support and have become increasingly popular (e.g., Babai, Eidelman, & Stavy, 2012; Dempster & Brainerd, 1995; De Neys & Everaerts, 2008; De Neys & Van Gelder, 2008; Houdé, 1997, 2000, 2007; Reyna, Lloyd, & Brainerd, 2003; Simoneau & Markovits, 2003). More specifically, there is also a rapidly growing field of literature on the importance of inhibition for mathematical learning (e.g., Attridge & Inglis, in press; Clayton & Gilmore, in press; Gillard, Van Dooren, Schaeken, & Verschaffel, 2009; Gilmore et al., 2013; Gilmore, Keeble, Richardson, & Cragg, in press; Lubin et al. 2013; Szücs, Devine, Soltesz, Nobes, & Gabriel, 2013; Van Hoof, Janssen, Verschaffel, & Van Dooren, 2014). However, the precise nature of children's inhibition failure when failing to solve arithmetic word problems is not clear. A key question is whether children fail the problems because they lack the executive resources to complete inhibiting the heuristic strategy or because they fail to detect that they need to inhibit the strategy in the first place. To clarify this point it is important to stress that inhibitory accounts do not posit that children always need to block their heuristic intuitions (e.g., Brainerd & Reyna, 2001; De Neys & Franssens, 2009; De Neys & Vanderputte, 2011; Houdé & Guichart, 2001; Jacobs & Klaczynski, 2002; Klaczynski, Byrnes, & Jacobs, 2001; Reyna et al., 2003; Stanovich, West, & Toplak, 2011). As we already noted, in many situations automatized heuristic strategies can provide a valid and useful basis for our judgment. Indeed, the no-conflict word problems that we introduced above are a very good illustration of this point. When the relational term does not conflict with the required mathematical operation, it is perfectly reasonable to rely on the heuristic. This implies that an efficient inhibition requires that one monitors for such conflict first and inhibits the heuristic strategy whenever it is detected. The detection might be quite implicit and boil down to a vague awareness that the heuristic response is not fully warranted (e.g., De Neys, 2012, 2014; Proulx, Inzlicht, & Harmon-Jones, 2012) but it is nevertheless a crucial building block for an efficient inhibition process. Hence, what we need to know is whether children show some minimal awareness of the questionability of their errors or not. Unfortunately, the efficiency of such an error detection process in simple arithmetic word problems has not been examined.

From a theoretical point of view, testing children's error detection skills is paramount to unravel the precise nature of their arithmetic failure. However, at a more applied level establishing whether or not children have some basic sensitivity with respect to their errors is also important to develop efficient intervention programs to *de-bias* their thinking. Existing general educational intervention programs aimed at reducing children's and adults' overreliance on heuristic impressions during reasoning have often focused on training participants' inhibitory processing capacities (e.g., Houdé, 2007; Houdé et al., 2000; Moutier, 2000; Moutier & Houdé, 2003). However, if younger children do not yet detect that the cued “add if more, subtract if less” heuristic is erroneous, such inhibition training will have less than optimal results in the case of arithmetic word problem solving. Clearly, any increase in inhibitory processing capacity per se is rather pointless if one is not able to determine whether or not it is needed to inhibit in the first place. Hence, examining children's error detection skills is paramount to determine which component an optimal intervention program needs to target.

In sum, both for theoretical and practical reasons it is important to test children's error detection efficiency during arithmetic word problem solving. In the present study we directly address this issue. We focused on the performance of a group of eight to eleven year-old elementary schoolchildren (third to fifth grade) because children in this age range are known to still have difficulties with arithmetic word problems (and we are obviously specifically interested in erroneous responses here, e.g., Lewis & Mayer, 1987; Morales et al., 1985; Riley et al., 1983). To test our hypothesis, children were given both conflict and no-conflict versions of simple arithmetic word problems. We therefore manipulated whether the relational term was consistent or inconsistent with the required mathematical operation.

After solving each version children were asked to indicate their response confidence on a simplified rating scale (e.g., De Neys & Feremans, 2013). This allowed us to measure children's error detection sensitivity. If erring elementary schoolchildren do not have an elementary understanding of the required mathematical operation or do not detect a conflict between their erroneous answer and this knowledge, their response confidence should not differ after solving conflict and no-conflict problems. However, if children have a minimal awareness of the unwarranted nature of their erroneous answer, this should decrease their confidence and result in lower confidence ratings after solving conflict than after solving no-conflict control problems.

2. Method

2.1. Participants

A total of 137 children (mean age ± SD = 9.8 ± .9 years, 67 girls) were tested. The children were recruited from three grades (grade 3, 4, and 5) from the same elementary school in Paris, France. We sampled an approximately even number of participants from each grade level (N = 41 for grade 3, N = 50 for grade 4, N = 46 for grade 5). All children had French as their mother tongue. The study was approved by the local school board and all parents or guardians gave informed consent for the study.

2.2. Material and procedure

Participants were collectively tested in their classroom. Booklets were distributed by the experimenter and children were instructed not to begin before receiving the general instruction. Before starting the experiment, the experimenter explained to the children that they had to solve four arithmetic problems. The children were then familiarized with the 5-point rating scale which was printed on the first page of the booklet. Five cylinders of different height were used to represent the different levels of the scale that ranged from 0 ("really not sure") to 4 ("totally sure"). The experimenter explained each point and gave some examples to check their scale understanding. Note that the training questions were designed to be either very easy ("what is your name?") or very hard ("Is Brussels a town in Suisse?"). The idea was that by presenting questions of varying difficulty we could quite naturally illustrate the response-confidence concept by pointing out that one might be sure about their response to some questions whereas one might be less sure about their response to other questions (e.g., De Neys, Lubin, & Houdé, 2014). For the hard questions, children were explicitly told by the experimenter that this was a really hard question and that they presumably felt unsure as to whether their answer was correct. Overall, the familiarization indicated that children had little trouble understanding the confidence question and rating scale. After the general instruction, the children were asked to write down their answer to each problem on the indicated space and then to circle the cylinder that best reflected their feeling of confidence (see Fig. 1).

Each participant was presented with a total of four problems in a booklet composed of 2 pages with 2 problems by page. Two of presented problems were conflict problems and the other two were no-conflict control problems. As illustrated in the introduction, the conflict and no-conflict version were created by manipulating whether the relational term was consistent or inconsistent with the required mathematical operation. An overview of all the problems can be found in the Appendix (Table 1). All four problems had a different content to keep children motivated and avoid potential cross-item interference. To minimize superficial content

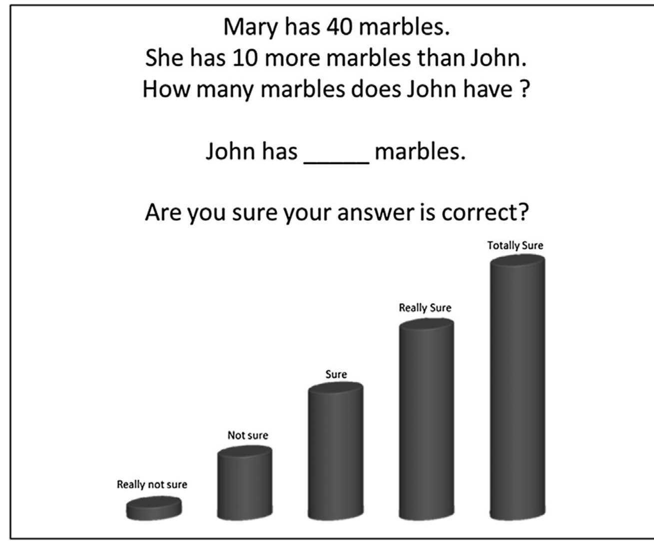

Fig. 1. An example of the item format with five-point confidence scale. Five cylinders were used to represent the different levels of the scale that ranged from 0 ("really not sure") to 4 ("totally sure"). The children had to write down their answer and to circle the cylinder that best reflected their feeling of confidence.

interference further we also used two slightly altered problem formats. Two problems (one conflict, one no-conflict) were based on the format used by Stern (1993) (e.g., There are 20 hens in the farm. There are 10 more hens than cows in the farm. How many cows are in the farm?). The two other problems (one conflict, one no-conflict) were based on the format used by Riley et al. (1983) (e.g., Mary has 40 marbles. She has 10 more marbles than John. How many marbles does John have?). These two formats have the same structural characteristics but the different format should make the task less repetitive for participants. We also counterbalanced the nature of the relational term that was used. For half of the participants all conflict and no-conflict problems used the relational term "more than", for the other half of the participants the relational term "less than" was used in all presented problems. Note that control analyses established that neither the format nor relational term affected the critical confidence findings (see results for details). Reported findings are averaged over these factors. For all problems we used simple operations (addition or subtraction) that involved numbers that were multiples of 10 (see Appendix, Table 1). Half of the presented booklets started with a conflict problem, while the other half started with a no-conflict problem.

One might note that the number of presented problems (i.e., four, two conflict and two no-conflict items) in the present study is somewhat smaller than in most studies on arithmetic word problem solving (e.g., Verschaffel, 1994). The rationale to limit the number of presented problems is based on a methodological concern in the work on error detection. Indeed, a key critique of error detection work in the reasoning and decision making field is that repeated testing with conflict and no-conflict items might lead to a cueing or implicit learning effect (De Neys, 2012; Kahneman, 2011). The concern is that in lengthy test sessions with multiple items the repeated alteration of conflict and no-conflict control items might artificially direct participants to start monitoring for conflict which might result in an overestimation of the error sensitivity. To sidestep this potential complication, we have opted to limit the number of items (e.g., see also De Neys, Rossi, & Houdé, 2013; De Neys et al., 2014; for a related approach). In a developmental context this has the additional advantage that the study is kept short and children have little trouble staying maximally motivated to answer the problems and indicate their confidence.

Each participant in the present study solved two conflict and two no-conflict problems. To get some overall indication of the reliability of the items we have calculated Cronbach's alpha for the accuracy and confidence measures. With respect to accuracy, Cronbach's alpha reached .69 for the conflict items (alpha was not computed for control problems given the ceiled performance and lack of variability). With respect to confidence ratings, Cronbach's alpha reached .56 for conflict problems and .70 for no-conflict problems. The somewhat lower value for conflict problems is not surprising given the less than perfect reliability of the conflict problem accuracy. If participants answer one problem correctly and incorrectly, the effect of accuracy (e.g., lower confidence for an incorrect vs. correct response) can attenuate the assessment of the confidence rating reliability per se. To address this problem we also calculated Cronbach alpha for the vast majority of reasoners ($n = 117$ out of 137) who answered both conflict problems consistently (i.e., two correct or incorrect conflict responses). This gives us a pure measure of the reliability of the conflict confidence rating per se. Results showed that in this corrected for attenuation analysis Cronbach alpha reached .70. By and large, this suggests that despite the limited number of items, our key measures have acceptable reliability.

3. Results and discussion

Accuracy. Consistent with previous findings, we observed that although our group of elementary schoolchildren performed fairly well overall, they had still some difficulties solving the conflict versions correctly (overall accuracy $\pm$ SE $= 78 \pm 3\%$). However, when there was no conflict and correct responding did not entail inhibition of the cued heuristic answer, children had little difficulty solving the task. Accuracy on the no-conflict control versions reached $99 \pm 1\%$ which was significantly better than the performance on the conflict versions, $F(1, 136) = 46.4$, $p < .00001$, $\eta p2 = .25$. In line with the expectations, the vast majority of the erroneous responses on the conflict problems (i.e., 85%) were reversal errors in which children added instead of subtracted (or vice versa).

Note that the virtually perfect performance on the no-conflict control problems indicates that participants' failure to solve the conflict problems cannot be attributed to a general lack of motivation or concentration confound. If children erred because they were not motivated to complete the task, their performance on the no-conflict control problems should have been equally affected.

Children in our study were recruited from three different grade levels. Although our study was not designed with a developmental goal in mind, for completeness, we also entered grade level as an additional factor in the accuracy analysis. This resulted in a 2 (Conflict, within-subjects) x 3 (Grade Level, between-subjects) mixed model ANOVA.[2] The analysis showed that there was a main effect of the Conflict factor, $F(1, 134) = 48.6$, $p < .00001$, $\eta p2 = .27$, a main effect of the Grade factor, $F(2, 134) = 4.2$, $p = .016$, $\eta p2 = .06$ and a significant interaction, $F(1, 80) = 3.5$, $p = .03$, $\eta p2 = .05$. Bonferroni post-hoc analyses established that performance on the conflict problems increased with grade (grade 5 > grade 4 = grade 3, $p < .05$). However, on the no-conflict control problems all grade levels performed at ceiling (all $p = 1$). These results are shown in Fig. 2.

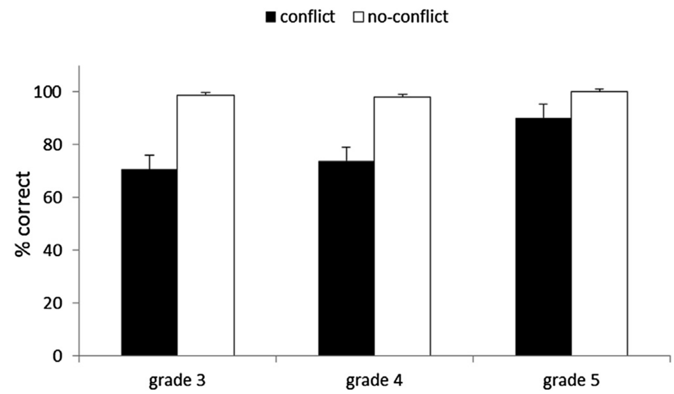

Fig. 2. Average percentage of correct responses on conflict and no-conflict control problems in different grades. Error bars are standard errors.

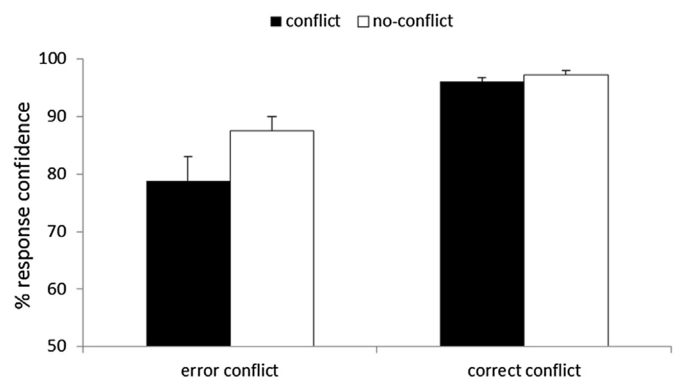

Fig. 3. Average response confidence (%) for erroneously and correctly solved conflict problems in contrast with the same individuals' average response confidence for correctly solved no-conflict problems. Error bars are standard errors.

Response confidence. The central question in our study concerned participants' response confidence. Obviously, to test the error detection sensitivity we were specifically interested in the confidence ratings for erroneously solved conflict problems. The key contrast concerns the confidence ratings for erroneously solved conflict problems and correctly solved no-conflict control problems. To recap, the no-conflict control problems can be solved correctly by mere reliance on the cued heuristic response. If children have a minimal awareness of the unwarranted nature of their erroneous answer when solving the conflict problems, this should decrease their confidence and result in lower confidence ratings after solving conflict than after solving no-conflict control problems. Therefore, for each individual we calculated the average confidence rating for erroneous conflict responses and correct no-conflict responses.[3] For ease of presentation all confidence ratings were rescored to percentage points. Results confirmed that there was a clear (about 9%) confidence drop for erroneous conflict responses, $F(1, 40) = 7.6$, $p < .01$, $\eta p2 = .16$. This is illustrated in Fig. 3. For completeness, we also ran the same contrast for correctly solved conflict problems (i.e., correct conflict confidence vs. correct

[2] To make sure that the basic dichotomous nature of the accuracy data was not distorting the findings the analysis was also run with non-parametric tests. Results were consistent. A Wilcoxon matched pairs test showed that there was a significant effect of conflict, $Z = -6.008$, $p < .001$. A Kruskal–Wallis test showed that the Grade effect was significant for conflict problems, $H(2, n = 137) = 7.687$, $p < .025$, but not for no-conflict problems, $H(2, n = 137) = 1.794$, $p = .408$.

[3] Reported analyses include all erroneous conflict responses. As noted, the majority of these errors were reversal errors. Note that we also ran control analyses that discarded the few non-reversal errors. Results were completely similar with the reported unrestricted results.

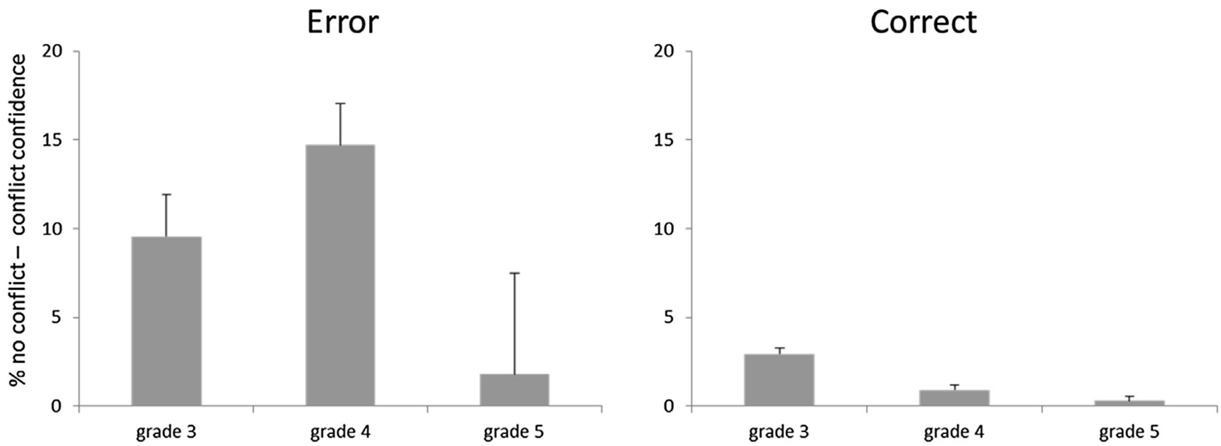

Fig. 4. Relative confidence difference scores (i.e., correct no-conflict confidence − conflict confidence) for erroneously (left panel) and correctly (right panel) solved conflict problems as a function of grade level. A higher value reflects a more pronounced confidence decrease when solving conflict problems. Error bars are standard errors.

no-conflict confidence). As one might expect, as Fig. 3 shows the confidence decrease was less pronounced (about 1%) for the correctly solved conflict problems, F (1, 118) = 3.88 p = .051, ηp2 = .03. In other words, when children manage to block the erroneous heuristic and answer the problem correctly, they also seem to realize that their answer is correct. A final contrast established that confidence in erroneously solved conflict problems was indeed significantly lower than confidence incorrectly solved conflict problems, F (1, 22) = 7.9, p = .01, ηp2 = .26.

To avoid confusion when inspecting Fig. 3, note that there were a considerable number of participants who did not make any errors on the conflict problems in the present study. Obviously, these individuals are not included in the incorrect conflict vs. correct no-conflict confidence contrast analysis. Fig. 3 indicates that in addition to a higher conflict confidence, these individuals also had a higher confidence in their answers on no-conflict problems.[4] However, the crucial finding is that despite the overall absolute lower confidence for the group of incorrect conflict responders, confidence was still lower for conflict than for no-conflict problems within this group. This directly establishes that incorrect conflict problem responders show sensitivity to their errors.

Finally, for completeness, we also entered grade level in the analysis for all above contrasts. There was always a significant effect of the Conflict factor (incorrect conflict vs correct no-conflict: F (1, 38) = 4.45, p < .05, ηp2 = .10; correct conflict vs. correct no-conflict: F (1, 116) = 4,63 p < .05, ηp2 = .04; incorrect conflict vs correct-conflict: F (1, 20) = 9.13, p < .01, ηp2 = .31), but the effect of Grade Level (all F < 1.92, all p > .16) and the Grade × Conflict interaction (all F < 1.69, all p > .21) never reached significance. Hence, by and large, the critical error sensitivity can be observed across the current age levels. Key results are illustrated in Fig. 4. For ease of interpretation the figure presents the relative confidence difference scores (i.e., correct no-conflict confidence − conflict confidence) for each contrast of interest.[5] A higher value reflects a more pronounced confidence decrease when solving conflict problems. As Fig. 4 indicates, if anything, the pattern for erroneous responses seemed to be slightly more pronounced for younger than for older reasoners. Hence, there is little evidence to suggest that the present results are driven by the oldest reasoners. Overall, older reasoners were less likely to be biased and to err, and in case they did err, they did not show a stronger confidence decrease than younger reasoners.

As we noted in the method section, to minimize superficial content interference and make the task less repetitive we presented each participant with two slightly altered problem formats. Two problems (one conflict, one no-conflict) were based on the format used by Stern (1993, study 3) whereas the other two (one conflict, one no-conflict) were based on the format used by Riley et al. (1983). In addition, we also counterbalanced the nature of the relational term that was used. For half of the participants all conflict and no-conflict problems used the relational term “more than”, for the other half of the participants the relational term “less than” was used in all presented problems. We had no a priori interest in these factors and the counterbalancing and crossing with the conflict factor guarantee that they cannot confound the findings. Nevertheless, for completeness one might want to verify whether these factors are affecting the observed error reversal sensitivity. Therefore, we entered the format (Stern vs. Riley content, within-subjects) and relational term (more vs. less, between-subjects) factors in our ANOVA (with Conflict and Grade level) on the confidence ratings for incorrectly solved conflict problems and correctly solved no-conflict problems. For the format factor this resulted in a 2 (format, within-subjects) x 2 (Conflict, within-subjects) x 3 (Grade, between-subjects) mixed model ANOVA, and for the relational term factor this also resulted in a 2 (relational term) x 2 (Conflict) x 3 (Grade) mixed model ANOVA. Key results for the relational term factor showed that there was a main effect of the relational term, F (1, 35) = 16.34, p < .001, ηp2 = .31, and a relational term by grade interaction, F (1, 35) = 6.23, p < .01, ηp2 = .26. Overall, confidence was higher when the relational term “less” was used and this was especially pronounced in the older grades. However, critically, there was a significant conflict effect, F (1, 35) = 6.83, p < .05, ηp2 = .16, and this factor did not interact with the relational term factor, F(1, 35) = 2.66, p = .11. The 3-way interaction between relational term, grade, and conflict was also not significant, F(1, 35) = 1.3, p = .28. With respect to the problem format analysis, the effect of conflict remained significant, F(1, 12) = 6.32, p < .05, ηp2 = .35, but neither the main effect of problem format, F(1, 12) < 1, nor its interactions with any of the other factors in the design reached significance, all Fs < 1. Taken together these analyses indicate that the conflict effect does not differ for each of the relational terms or formats that were used. To be clear, it should be noted that the format factor analysis is run on a small number of participants (n = 15) and with only one item per

[4] A between subject analysis that contrasted the no-conflict response confidence for the subgroups of participants who solved all or none of the conflict problems correctly suggested this was indeed the case, F (1, 110) = 21.7, p < .0001, ηp2 = .16).

[5] A full table with all absolute values can be found in the Appendix.

condition (i.e., since we are interested in reversal errors we can only include participants who err on both conflict problems to assess the error sensitivity). Hence, caution is needed when interpreting these analyses.

Confidence and accuracy correlation. Each participant in the current study was presented with two conflict problems. One might wonder whether a higher accuracy rate among biased reasoners (i.e., whether one erred on both or only one conflict problem) was associated with a stronger error sensitivity effect (i.e., a stronger confidence decrease; e.g., Mevel et al., 2015). We therefore calculated the correlation between a participants' accuracy on the conflict problems and the size of the confidence decrease (i.e., average confidence for correct no-conflict responses - average confidence for erroneously solved conflict problems). Although there was a trend towards a positive association, the effect failed to reach significance, Spearman $R = .22$, $p = .17$. *A* 2 (Conflict, conflict vs. no-conflict) x 2 (Bias level, one vs. two errors) mixed model ANOVA in which conflict accuracy (i.e., bias level) was entered as a between subjects factor confirmed that neither the main effect of accuracy, $F(1, 39) < 1$, nor its interaction with the conflict factor, $F(1, 39) = 2.76$, $p > .10$, reached significance. In line with the grade level findings this suggests that more and less biased reasoners alike show sensitivity to their errors.

4. General discussion

The present findings establish that in those cases where elementary schoolchildren fail to solve simple arithmetic word problems, they show sensitivity to their errors. Although they might fail to provide a correct answer and succumb to the "add if more/subtract if less" heuristic, their decreased response confidence indicates that they at least detect that their response is not fully warranted. These data directly argue against the view that these errors result from a lack of mathematical knowledge or mathematical sophistication per se. If erring elementary schoolchildren do not have an elementary understanding of the required mathematical operation or do not detect a conflict between their erroneous answer and this mathematical knowledge, they should have no reason to doubt their answer.

It might be interesting to link the present work on arithmetic word problems to research on bias detection during logical and probabilistic thinking in the reasoning and decision-making field. Classic studies on reasoning and decision-making have long established that people's inferences are often biased by prior beliefs and stereotypical intuitions (e.g., Evans, 2010; Kahneman & Tversky, 1973). In line with the present findings, it has recently been shown that reasoners also detect the biased nature of their intuitive logical and probabilistic judgments (e.g., Bonner & Newell, 2010; De Neys, 2012; De Neys & Bonnefon, 2013; De Neys et al., 2013; Mevel et al., 2015; Morsanyi & Handley, 2012; Pennycook & Thompson, 2012; Stupple & Ball, 2008; Villejoubert, 2009). Interestingly, however, developmental studies have suggested that this bias detection during logical and probabilistic reasoning is only observed after the onset of adolescence (i.e., by the end of elementary school, e.g., De Neys, Cromheeke, & Osman, 2011; De Neys & Feremans, 2013). This developmental pattern has been linked to the late maturation of the Anterior Cingulate Cortex (ACC), the critical brain structure that is supposed to be mediating conflict and error detection, which only achieves full functionality over the adolescent years (e.g., Davies, Segalowitz, & Gavin, 2004; De Neys, Vartanian, & Goel, 2008; Fitzgerald et al., 2010; Santesso & Segalowitz, 2008). Given these findings, the presently established successful arithmetic error detection in our slightly younger sample of third to fifth graders might seem somewhat surprising. Indeed, our developmental analysis failed to detect an effect of grade level and indicates that the error detection effect can be observed even among biased eight-year-olds. Clearly, our somewhat restricted age range and the fact that error rates were very low in the oldest grade imply that these fine-grained age trends (or lack thereof) need to be interpreted with some caution. However, it is also important to take into account that a less developed ACC does not imply a lack of all conflict detection (Lubin, Simon, Houdé, & De Neys, in press). Indeed, basic error monitoring studies have shown that even three-year-olds can detect errors in simple tasks that do not cue a strong intuitive response (Lyons & Ghetti, 2011). Likewise, De Neys et al. (2014) showed that preschoolers can show sensitivity to errors on Piaget's classic number conservation tasks. Arguably, in comparison with logical and probabilistic reasoning tasks in which the cued intuitive response typically entails a semantic prior belief or stereotypical information, the "add if more/subtract if less" heuristic might be less tempting and easier to monitor (and subsequently inhibited). This would also account for the relatively high level of correct responses in the present study (e.g., 78%) in comparison with logical and probabilistic reasoning tasks where (even for adults) correct performance typically only hovers around 20% (e.g., De Neys et al. 2011; De Neys & Feremans, 2013; Kahneman, 2011). Hence, detection of heuristic bias in simple arithmetic word problems might be less demanding and occur at a younger age than in logical and probabilistic reasoning tasks.

From a methodological point of view one might wonder whether it could be interesting to ask children to give an explicit justification of their confidence rating. It is important to note here that the work on conflict detection in the reasoning and decision making field has indicated that conflict detection is typically implicit in nature (e.g., De Neys & Glumicic, 2008; De Neys & Vanderputte, 2011; Proulx et al., 2012; see De Neys, 2014; for review). People can detect that the heuristic response is erroneous but even adults do not necessarily manage to verbally explicitate why this is the case (e.g., De Neys & Glumicic, 2008). Hence, although it can be interesting to see whether children can explain confidence scores, it will not necessarily be the case that error sensitivity will be reflected in participants' explicit verbalizations. In that sense, the explanations do not give us a critical test of the error sensitivity hypothesis. The basic confidence ratings are more informative here.

We noted that the present study has also pedagogical implications with respect to the design of intervention programs. For example, the evidence for children's arithmetic word problem error sensitivity indicates that there is little point in running programs that focus on a familiarization and teaching of the required arithmetical principles per se. Bluntly put, if the problem is not that children do not know the required operation, merely informing them about it will not be very helpful. Rather, a more promising approach seems to be to focus on training children's capacities to override their erroneous heuristic intuitions (Lubin et al., in press). As we noted, existing inhibitory training programs have been shown to be successful at reducing children's and adults' overreliance on intuitive impressions during reasoning and decision-making in the lab (e.g., Houdé, 2007; Houdé et al., 2000; Moutier, 2000; Moutier, Angeard, & Houdé, 2002; Moutier & Houdé, 2003) and in school (Lubin, Lanoë, Pineau, & Rossi, 2012; Rossi, Lubin, Lanoë, & Pineau, 2012). Since the present evidence suggests that third to fifth graders can reliably distinguish between situations in which their heuristic intuitions violate the required mathematical operations when solving arithmetic word problems or not, such inhibitory training programs might prove to be highly efficient to boost these children's accuracy rates.

At the same time, we like to stress that although it is undeniably important to get children to respond correctly to arithmetic word problems, our confidence findings underscore the dangers of a mere reliance on response output (i.e., accuracy) to measure children's mathematical sophistication (e.g., Verschaffel, 1994; Zelazo & Müller, 2011). A simple behavioral process measure such as one's response confidence indicates that erring third to fifth graders have a better arithmetic understanding than their mere errors might seem to suggest. Indeed, the simple fact that erring children detect that their answer is questionable when it conflicts with the required mathematical operation implies that they are more knowledgeable than their test accuracy indicates.

Acknowledgments

Preparation of this manuscript was supported by a grant from the French National Research Agency (ANR-12-JSH2-0007-01). We thank Mostafa Quenum-Sanfo for his valuable help with the data collection. We thank the children and parents who took part in this experiment, their teachers and the School Inspectorate.

Appendix

Table 1
Overview of all the presented problems

Conflict problem	Format 1. Mary has 40 marbles. She has 10 more (less) marbles than John. How many marbles does John have? Format 2. There are 20 hens in the farm. There are 10 more (less) hens than cows in the farm. How many cows are in the farm?
No conflict control problem	Format 1. Jane has 50 apples. Sarah has 20 more (less) apples than Jane. How many apples does Sarah have? Format 2. There are 30 lions in the zoo. There are 20 more (less) tigers than lions in the zoo. How many tigers are in the zoo?

Table 2
Average response confidence (%) for erroneously and correctly solved conflict problems in contrast with the same individuals' average response confidence for correctly solved no-conflict problems in different grades.

	Grade	Conflict confidence (±SE)	No-conflict confidence (±SE)
Incorrect conflict responses	Grade 3	85.3 ± 6%	94.9 ± 3.6%
	Grade 4	76.5 ± 6%	91.2 ± 3.6%
	Grade 5	75 ± 3.6%	76.8 ± 9.3%
Correct conflict responses	Grade 3	94.9 ± 1.7%	97.8 ± 1.4%
	Grade 4	96.3 ± 1.6%	97.3 ± 1.3%
	Grade 5	96.6 ± 1.5%	96.9 ± 1.2%

References

Attridge, N., & Inglis, M. Increasing cognitive inhibition with a difficult prior task: implications for mathematical thinking. *ZDM – The International Journal on Mathematics Education*. http://dx.doi.org/10.1007/s11858-014-0656-1 (in press).

Babai, R., Eidelman, R. R., & Stavy, R. (2012). Preactivation of inhibitory control mechanisms hinders intuitive reasoning. *International Journal of Science and Mathematics Education, 10*, 763–775. http://dx.doi.org/10.1007/s10763-011-9287-y.

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition, 38*, 186–196.

Borst, G., Moutier, S., & Houdé, O. (2013). Negative priming in logicomathematical reasoning: the cost of blocking your intuition. In W. De Neys, & M. Osman (Eds.), *New approaches in reasoning research – Current issues in thinking & reasoning*. New York: Psychology Press.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory : dual processes in memory, reasoning, and cognitive neuroscience. In H. W. Reese, & R. Kail (Eds.), *Advances in child development and behavior* (vol. 28, pp. 41–100). San Diego: Academic Press.

Clayton, S., & Gilmore, C. Inhibition in dot comparison tasks. *ZDM – The International Journal on Mathematics Education*. http://dx.doi.org/10.1007/s11858-014-0655-2 (in press).

Davies, P. L., Segalowitz, S. J., & Gavin, W. J. (2004). Development of response-monitoring ERPs in 7-to 25-year-olds. *Developmental Neuropsychology, 25*, 355–376.

De Corte, E., & Verschaffel, L. (1987). The effect of semantic structure on first graders' strategies for solving addition and subtraction word problems. *Journal for Research in Mathematics Education, 18*, 363–381.

De Corte, E., Verschaffel, L., & Pauwels, A. (1990). Influence of the semantic structure of word problems on second graders' eye movements. *Journal of Educational Psychology, 82*(2), 359–365.

De Neys, W. (2012). Bias and conflict: a case for logical intuitions. *Perspectives on Psychological Science, 7*, 28–38.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: some clarifications. *Thinking & Reasoning, 20*, 169–187.

De Neys, W., & Bonnefon, J. F. (2013). The whys and whens of individual differences in thinking biases. *Trends in Cognitive Sciences, 17*, 172–178.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: conflict and decision confidence. *PLoS One, 6*, e15954. http://dx.doi.org/10.1371/journal.pone.0015954.

De Neys, W., & Everaerts, D. (2008). Developmental trends in everyday conditional reasoning: the retrieval and inhibition interplay. *Journal of Experimental Child Psychology, 100*, 252–263.

De Neys, W., & Feremans, V. (2013). Development of heuristic bias detection in elementary school. *Developmental Psychology, 49*, 258–269.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: not always winning but at least taking part. *Cognition, 113*, 45–61.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of reasoning. *Cognition, 106*, 1248–1299.

De Neys, W., Lubin, A., & Houdé, O. (2014). The smart non-conserver: preschoolers detect their number conservation errors. *Child Development Research*. http://dx.doi.org/10.1155/2014/768186, 2014, Article ID 768186, 7.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20*, 269–273.

De Neys, W., & Van Gelder, E. (2008). Logic and belief across the life span: the rise and fall of belief inhibition during syllogistic reasoning. *Developmental Science, 12*, 123–130.

De Neys, W., & Vanderputte, K. (2011). When less is not always more: stereotype knowledge and reasoning development. *Developmental Psychology, 47*, 432–441.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: when our brains detect that we are biased. *Psychological Science, 19*, 483–489.

Dempster, F. N., & Brainerd, C. J. (1995). *Interference and inhibition in cognition*. San Diego, CA: Academic press.

Evans, J. B. S. T. (2010). Intuition and reasoning: a dual process perspective. *Psychological Inquiry, 21*, 313–326.

Fitzgerald, K. D., Perkins, S. C., Angstadt, M., Johnson, T., Stern, E. R., Welsh, R. C., et al. (2010). The development of performance-monitoring function in the posterior medial frontal cortex. *Neuroimage, 49*, 3463–3473.

Gillard, E., Van Dooren, W., Schaeken, W., & Verschaffel, L. (2009). Dual processes in the psychology of mathematics education and cognitive psychology. *Human Development, 52*, 95–108.

Gilmore, C., Attridge, N. F., Clayton, S., Cragg, L., Johnson, S., Marlow, N., et al. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. *PLoS One, 8*, e67374. http://dx.doi.org/10.1371/journal.pone.0067374.

Gilmore, C., Keeble, S., Richardson, S., & Cragg, L. The role of cognitive inhibition in different components of arithmetic. *ZDM – The International Journal on Mathematics Education*. http://dx.doi.org/10.1007/s11858-014-0659-y (in press).

Giroux, J., & Ste-Marie, A. (2001). The solution of compare problems among first-grade students. *European Journal of Psychology of Education, 16*, 141–161.

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: evidence from students' eye fixations. *Journal of Educational Psychology, 84*, 76–84.

Houdé, O. (1997). Rationality in reasoning: the problem of deductive competence and the inhibitory control of cognition. *Current Psychology of Cognition, 16*, 108–113.

Houdé, O. (2000). Inhibition and cognitive development: object, number, categorization, and reasoning. *Cognitive Development, 15*, 63–73.

Houdé, O. (2007). First insights on neuropedagogy of reasoning. *Thinking & Reasoning, 13*, 81–89.

Houdé, O., & Guichart, E. (2001). Negative priming effect after inhibition of number/length interference in a Piaget-like task. *Developmental Science, 4*, 71–74.

Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., et al. (2000). Shifting from the perceptual brain to the logical brain: the neural impact of cognitive inhibition training. *Journal of Cognitive Neuroscience, 12*, 721–728.

Jacobs, J. E., & Klaczynski, P. A. (2002). The development of decision making during childhood and adolescence. *Current Directions in Psychological Science, 4*, 145–149.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, & Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Klaczynski, P. A., Byrnes, J. B., & Jacobs, J. E. (2001). Introduction: special issue on decision making. *Journal of Applied Developmental Psychology, 22*, 225–236.

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79*, 363–371.

Lubin, A., Lanoë, C., Pineau, A., & Rossi, S. (2012). Inhibition training: an innovative pedagogy for academic learning (mathematics and spelling) in schoolchildren aged 6 to 11 years (title in French: Apprendre à inhiber: une pédagogie innovante au service des apprentissages scolaires fondamentaux (mathématiques et orthographe) chez des élèves de 6 à 11 ans). *Neuroéducation, 1*, 55–84.

Lubin, A., Poirel, N., Rossi, S., Pineau, A., & Houdé, O. (2009). Math in actions: actor mode reveals the true arithmetic abilities of French-speaking 2-year-olds in a magic task. *Journal of Experimental Child Psychology, 103*, 376–385.

Lubin, A., Simon, G., Houdé, O., & De Neys, W. Inhibition, conflict detection, and number conservation. *ZDM – The International Journal on Mathematics Education*. http://dx.doi.org/10.1007/s11858-014-0649-0 (in press).

Lubin, A., Vidal, J., Lanoë, C., Houdé, O., & Borst, G. (2013). Inhibitory control is needed for the resolution of arithmetic word problems: a developmental negative priming study. *Journal of Educational Psychology, 105*, 701–708.

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development, 82*, 1778–1787.

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., et al. (2015). Bias detection: response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology, 27*, 227–237. http://dx.doi.org/10.1080/20445911.2014.986487.

Morales, R. V., Shute, V. J., & Pellegrino, J. W. (1985). Developmental differences in understanding and solving simple mathematics word problems. *Cognition and Instruction, 2*, 41–57.

Morsanyi, K., & Handley, S. (2012). Does thinking make you biased? the case of the engineers and lawyer problem. *Proceedings of the Annual Meeting of the Cognitive Science Society, 34*, 2049–2054.

Moutier, S. (2000). Deductive reasoning and experimental matching-bias inhibition training in school children. *Current Psychology of Cognition, 19*, 429–452.

Moutier, S., Angeard, N., & Houdé, O. (2002). Deductive reasoning and matching-bias inhibition training : evidence from a debiasing paradigm. *Thinking and Reasoning, 8*, 205–224.

Moutier, S., & Houdé, O. (2003). Judgment under uncertainty and conjunction fallacy inhibition training. *Thinking and Reasoning, 9*, 185–201.

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base-rates is routine, relatively effortless and context-dependent. *Psychonomic Bulletin & Review, 19*, 528–534.

Proulx, T., Inzlicht, M., & Harmon-Jones, E. (2012). Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in Cognitive Sciences, 16*, 285–291.

Reyna, V. F., Lloyd, F. J., & Brainerd, C. J. (2003). Memory, development, and rationality: an integrative theory of judgement and decision-making. In S. Schneider, & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 201–245). New York: Cambridge University Press.

Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5*, 49–101.

Riley, M. S., Greeno, J. G., & Heller, J. H. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsberg (Ed.), *the development of mathematical thinking* (pp. 153–196). San Diego, CA: Academic Press.

Rossi, S., Lubin, A., Lanoë, C., & Pineau, A. (2012). A pedagogical approach of cognitive control to improve attention to instructions in preschool children (title in French: une pédagogie du contrôle cognitif pour l'amélioration de l'attention à la consigne chez l'enfant de 4-5 ans). *Neuroéducation, 1*, 29–54.

Santesso, D. L., & Segalowitz, S. J. (2008). Developmental differences in error-related ERPs in middle- to late-adolescent males. *Developmental Psychology, 44*, 205–217.

Schumacher, R. F., & Fuchs, L. S. (2012). Does understanding relational terminology mediate effects of intervention on compare word problems? *Journal of Experimental Child Psychology, 111*, 607–628.

Simoneau, M., & Markovits, H. (2003). Reasoning with premises that are not empirically true: evidence for the role of inhibition and retrieval. *Developmental Psychology, 39*, 964–975.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). The complexity of developmental predictions from dual process models. *Developmental Review, 31*, 103–118.

Stern, E. (1993). What makes certain arithmetic word problems involving the comparison of sets so difficult for children? *Journal of Educational Psychology, 85*, 7–23.

Stern, E., & Lehrndorfer, A. (1992). The role of situational context in solving word problems. *Cognitive Development, 7*, 259–268.

Stupple, E. J. N., & Ball, L. J. (2008). Belief-logic conflict resolution in syllogistic reasoning: inspection-time evidence for a parallel-process model. *Thinking & Reasoning, 14*, 168–181.

Szücs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2013). Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. *Cortex, 49*, 2674–2688.

Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology, 37A*, 571–590.

Tirosh, D., Tsamir, P., & Hershkovitz, S. (2008). Insights into children's intuitions of addition, subtraction, multiplication, and division. In A. Cockburn, & G. Littler (Eds.), *Mathematical misconceptions*. , New Delhi: Sage Publications.

Vamvakoussi, X., Dooren, W. V., & Verschaffel, L. (2012). Brief Report. Educated adults are still affected by intuitions about the effect of arithmetical operations: evidence from a reaction-time study. *Educational Studies in Mathematics, 82*, 323–330.

Van Hoof, J. V., Janssen, R., Verschaffel, L., & Dooren, W. V. (2014). Inhibiting natural knowledge in fourth graders: towards a comprehensive test instrument. *ZDM Mathematics Education*, 1–9. http://dx.doi.org/10.1007/s11858-014-0650-7.

Verschaffel, L. (1994). Using retelling data to study elementary-school children representations and solutions of compare problems. *Journal for Research in Mathematics Education, 25*, 141–165.

Verschaffel, L., de Corte, E., & Pauwels, A. (1992). Solving compare problems: an eye movement test of Lewis and Mayer's consistency hypothesis. *Journal of Educational Psychology, 84*, 85–94.

Villejoubert, G. (2009). Are representativeness judgments automatic and rapid? the effect of time pressure on the conjunction fallacy. *Proceedings of the Annual Meeting of the Cognitive Science Society, 30*, 2980–2985.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358*, 749–750.

Zelazo, P. D., & Müller, U. (2011). Executive function in typical and atypical development. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development: Second edition* (pp. 574–603). Blackwell Publishing.