Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-019-01628-3

When masters of abstraction run into a concrete wall: Experts failing arithmetic word problems

Hippolyte Gros [1,2] · Emmanuel Sander [2] · Jean-Pierre Thibaut [3]

Abstract
Can our knowledge about apples, cars, or smurfs hinder our ability to solve mathematical problems involving these entities? We argue that such daily-life knowledge interferes with arithmetic word problem solving, to the extent that experts can be led to failure in problems involving trivial mathematical notions. We created problems evoking different aspects of our non-mathematical, general knowledge. They were solvable by one single subtraction involving small quantities, such as 14 – 2 = 12. A first experiment studied how university-educated adults dealt with seemingly simple arithmetic problems evoking knowledge that was either congruent or incongruent with the problems' solving procedure. Results showed that in the latter case, the proportion of participants incorrectly deeming the problems "unsolvable" increased significantly, as did response times for correct answers. A second experiment showed that expert mathematicians were also subject to this bias. These results demonstrate that irrelevant non-mathematical knowledge interferes with the identification of basic, single-step solutions to arithmetic word problems, even among experts who have supposedly mastered abstract, context-independent reasoning.

Keywords Encoding effects · Mathematical cognition · Mental models · Semantics

Introduction

Is 14 − 2 = 12 always obvious? Most third graders know the basics of addition and subtraction (Carpenter & Moser, 1984), and solving elementary arithmetic operations is no big deal from this point onwards. We learn from an early age that operations such as 14 – 2 = 12 are always valid, no matter whether one is subtracting apples, cars, or smurfs. However, our claim is that adults whose mathematical knowledge is unquestionable, even outstanding, sometimes fail to solve arithmetic problems involving a single-step solution such as 14 – 2 = 12 when their knowledge about the entities subtracted interferes with the mathematical structure of the problem.

✉ Hippolyte Gros
hippolyte.gros@cri-paris.org

[1] Center for Research and Interdisciplinarity, Paris Descartes University, Paris, France

[2] IDEA Lab, Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

[3] Lead, CNRS UMR 5022, University of Bourgogne Franche-Comté, Bourgogne, France

This prediction arises from a growing body of literature suggesting that the daily-life, non-mathematical world knowledge one has about the objects an arithmetic word problem refers to might influence their mathematical representation of the problem and their subsequent choice of a solving strategy. For example, Bassok, Wu, and Olseth (1995) showed that being trained to solve a permutation problem was not always helpful to solve analogous problems. The authors demonstrated that slight, mathematically-irrelevant changes in the semantic relations linking the objects mentioned in the cover stories (e.g., computers assigned to secretaries vs. secretaries assigned to computers) led to significant performance differences. Subsequent research has shown that non-mathematical semantic information related to the entities described in a problem influences lay solvers' performance (Bassok, Chase, & Martin, 1998; Gros, Sander, & Thibaut, 2016; Thevenot & Barrouillet, 2015; Verschaffel, De Corte, & Vierstraete, 1999; Vicente, Orrantia, & Verschaffel, 2007) as well as strategy choice (Gamo, Sander, & Richard, 2010; Gros, Thibaut, & Sander, 2017) and transfer (Gros, Thibaut, & Sander, 2015) on arithmetic word problems. Most of the available evidence regarding this issue has been collected with children and non-expert adults on problems that were not straightforward (e.g., complex permutation problems). Building on this literature, we propose to go further and show

that irrelevant aspects of what we call *world semantics* (the non-mathematical knowledge about the world that is evoked by the entities described in a specific problem statement) can also mislead experts in mathematics on problems involving basic arithmetic notions, despite them being considered experts in abstract, context-independent reasoning (Dehaene, 2011). We call this proposal the "world semantics view."

Despite stemming from the aforementioned literature, the claim that world semantics could exert such a pervasive influence and threaten even the highest levels of mathematical expertise is rather innovative, as it challenges the commonly held view in the expertise literature regarding experts' proficiencies. This expertise view notably considers that experts identify what has been described as the "deep structure" of the problem (Chi, Feltovich, & Glaser, 1981), its "principle" (Ross, 1987), its "objective mathematical structure" (Bassok, 2001), or its "problem space" (Newell & Simon, 1972). This deep structure is independent of the semantics imbued in the problem statement, and as such it is the foundation of experts' abstract, context-independent reasoning about the problem. Indeed, since by definition mathematics is not empirical and manipulates abstract symbols rather than real-life objects (Davis, Hersh, & Marchisotto, 2011; Russell, 1903), mathematical experts should ignore irrelevant information associated with the entities on which numbers and algorithms operate. They should perceive the deep structure of arithmetic problems that can be solved by simple subtractions (i.e., involving small quantities such as 14 − 2), no matter whether they calculate the price of an apple, the height of a smurf, or the speed of a car. Furthermore, experts are known to show exceptional performance in domain-related tasks (Chi, 2006), they stand out in their ability to generate problem solutions (De Groot, 1965), to detect relevant problem features (Lesgold et al., 1988), to monitor their own comprehension (Chi, 1978), and to qualitatively analyze the task at hand (Voss, Greene, Post, & Penner, 1983) (see Chi, 2006 for a review of experts' proficiencies). These former studies do not predict that the semantics conveyed by the problem statement could interfere with the experts' understanding of the problems' mathematical structure.

We performed two experiments to show that, contrary to this expertise view – but in accordance with the world semantics view – arithmetic problems admitting a single-step solution might pose a challenge to mathematical experts. We presented participants with a series of isomorphic problems involving two numerical values. Crucially, for each problem, a solution was provided (a single subtraction between the problem's two numerical values), and participants' task was to evaluate its validity. By varying the semantic, non-mathematical information evoked by the problem statements (e.g., use of an elevator vs. a weighing scale, reference to marbles being won vs. years passing by, mention of hamburger prices vs. statues' heights, etc.), we intended to show that even math experts are exposed to a deleterious influence of the non-mathematical knowledge evoked by the problem statement.

Our world semantics view predicts that university students (Study 1) – and math experts (Study 2) – will more often fail to recognize the proposed solution when it conflicts with the non-mathematical knowledge about the world evoked by the entities featured in the problem statement than when the solution is consistent with it. Furthermore, it predicts that a recoding process, akin to re-representation (Davidson & Sternberg, 2003; Vicente et al., 2007) is necessary when a problem's initial encoding leads to a dead end. Indeed, when the semantic content of a problem statement leads participants to interpret the situation in a way that is not compatible with the problem's solution, then it becomes necessary to build a new representation of the situation congruent with the solution. When successfully performed, such a recoding process should result in longer response times for correct answers conflicting with the problems' world semantics.

Study 1

Methods

Participants

We recruited 85 adults (50 women, mean age = 23.35 years, *SD* = 7.82) in the Paris region. All had attended university (mean length of university curriculum = 2.85 years, *SD* = 1.18), but none majored in mathematics. Considering the low complexity of the math problems involved, participants' curriculum was a clear indicator that they possessed the mathematical expertise required to solve the problems. Sample size was determined using uncertainty and publication bias correction on results from a previous study (Gros et al., 2016), following Anderson, Kelley, and Maxwell's recommendations (2017).

Materials

Our materials were inspired by Gamo et al. (2010), who showed that problems with the same formal mathematical structure are nevertheless preferentially solved with one of two available solving strategies, depending on the semantic content of the problem. Consider the weight problem in Table 1: this problem can be solved through two strategies. One is a three-step algorithm consisting of calculating the weight of each individual dictionary to compute the weight of the stack of dictionaries Lola is carrying: 14 – 5 = 9; 5 – 2 = 3; 9 + 3 = 12. The other one is a one-step algorithm that requires understanding that since Lola and Joe carry the same Spanish dictionary, calculating the weight of each book is

Table 1 Two isomorphic problems sharing the same mathematical structure but evoking different aspects of our knowledge about the world

Weight problem	Duration problem
Joe takes a Russian dictionary weighing 5 kg	Tom took painting classes for 5 years
He also takes a Spanish dictionary	He started taking painting classes at a specific age
In total, he is carrying 14 kg of books	He stopped taking the classes at the age of 14 years
Lola takes Joe's Spanish dictionary and a German dictionary	Lucy started taking painting classes at the same age as Tom
The German dictionary weighs 2 kg less than the Russian dictionary	She took classes for 2 years less than him
How many kilograms of books is Lola carrying?	How old was Lucy when she stopped taking painting classes?

unnecessary. Since the German dictionary is 2 kg lighter than the Russian dictionary, the weight difference between Joe's and Lola's books is of 2 kg as well: 14 – 2 = 12.

The duration problem in Table 1 has the same mathematical structure and can be solved using the same solving procedures. However, Gamo et al. (2010) showed that the two solving procedures are not randomly distributed across the two types of problems. Participants favor the three-step algorithm on problems like the dictionary problem (called *cardinal* problems) and the one-step algorithm on the second type of problems (called *ordinal* problems). This strategy using imbalance was our starting point. Gamo et al. (2010) and Gros et al. (2017) showed that the differences in the world semantics evoked by the problems resulted in different spontaneous encodings of the situations, from which this imbalance originated[1] (see Fig. 1 for a description of this effect). Since cardinal and ordinal problems shared the same structure featuring the same parts and wholes presented in the same order with the same numerical values, the imbalance in strategy use could only be attributed to the variations of the semantic content of the problem statements. Additionally, when considering the correct answers on either algorithm there was no significant difference in adults' performance between cardinal and ordinal problems, which indicates that the strategy imbalance was not a matter of problem difficulty (Gros et al., 2017).

Gros et al. (2017) have shown that most adults encode collection, price, and weight problems as cardinal representations, whereas they encode duration, distance, and floor problems as ordinal representations. We modified their problems and removed the value of *Part 1* so that the three-step strategy could not be used (see Table 2). Consequently, the only solution left was the one-step strategy, which required using the values of *Whole 1* and of the *Difference* (see Fig. 1). The constructed materials are available online (https://osf.io/fxgqh/?view_only=ed1374ef4d204c90a0cb03a30cb0a099). Ordinal problems were 333.5 characters long on average (SD = 38.37) and cardinal problems were 304 characters long on average (SD = 44.94). This length difference was not statistically significant ($t(10) = 1.18$, $p = .26$, paired t-test). Crucially, for each problem, participants were presented with the correct one-step solution (e.g. "14 – 2 = 12; Jolene has 12 marbles"). Participants' task was to decide whether the provided solution worked, or whether there was no solution to the problem. Due to the already established imbalance in strategy use between problems evoking a cardinal encoding and problems evoking an ordinal encoding (Gamo et al., 2010; Gros et al., 2017), we assumed that the measure of participants' ability to use the only remaining strategy on problems evoking different aspects of world semantics would be an effective assessment of the robustness of these effects.

The world semantics hypothesis predicts lower performances on cardinal than on ordinal problems, even among experts, because cardinal problems would require a re-representation of the situation when the only solution available is the one-step algorithm. By contrast, ordinal problems should be easier to solve because participants' spontaneous encoding facilitates the use of the one-step algorithm. Since university-educated adults can be considered experts in solving subtractions such as 14 2 = 12, and since the deep structure of a problem is identical regardless of the objects involved, this prediction could not be made without the world semantics view, especially when participants only need to check the validity of the proposed solution. Additionally, we predict that recoding a situation initially encoded as a combination of subsets (such as a cardinal encoding) into a representation in terms of states and transitions between states (such as an ordinal encoding) is a costly process, requiring a longer response time. Although our hypotheses only regard solvable problems, we also included unsolvable distractors in the materials, so that the correct answer would not always be "This problem can be solved." Among those distractors the

[1] Although the explanation of this effect is not the purpose of the present paper, the authors suggest that because our world knowledge about dictionaries says we can stack them with no specific order, they evoke a representation of the total as a combination of subsets, which they call a cardinal representation. A similar reasoning can be held for weights or prices defined as object properties (Gros et al., 2017). On the other hand, using the one-step algorithm requires participants to build a re-representation of the problem that is not based on a "combination of subsets," which makes computing the weight of the Spanish dictionary unnecessary. By contrast, some problems seem to emphasize the ordinal nature of the values featured and afford a representation of the numerical values on a continuous axis. For example, we spontaneously encode durations on a timeline, which makes it easier for school children and lay adults to notice that the numerical difference between the two distinct parts is equal to the difference between the two totals (Gamo et al., 2010). A similar reasoning can be held for height or floor problems (Gros et al., 2017). Thus, using the one-step algorithm is more straightforward for ordinal than for cardinal problems (see Fig. 1)

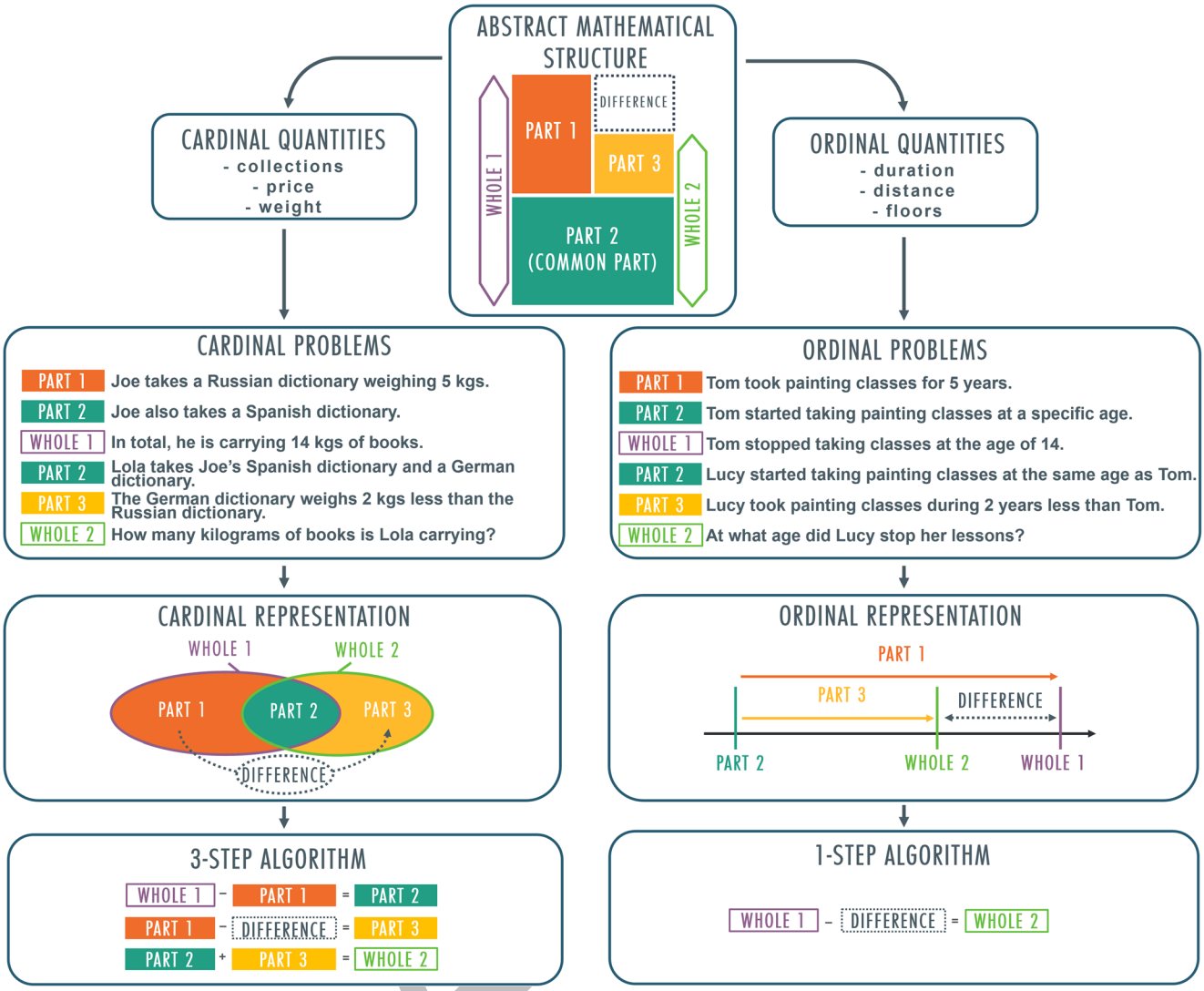

Fig. 1 Implementation of the mathematical structure with ordinal versus cardinal quantities, leading to different problem statements, representations, and strategy use

value of *Whole 1* was removed instead of the value of *Part 1*, which rendered the problems unsolvable with either algorithm.

Procedure

Participants answered the questions using three keyboard keys on a 17-in. laptop. Instructions stated that "Some of the problems can be solved using the values provided, while other problems cannot be solved with the available information. Your task is to tell apart problems that can be solved from problems that cannot. Answer as quickly as you can, although being correct is more important than being fast."

Participants were presented with six target problems that were only solvable with the one-step algorithm: three cardinal and three ordinal problems. An equal number of distractors was introduced to fulfill subjects' expectations regarding the uniform distribution of yes/no answers. Problem order, cover stories, and numerical values were randomized between participants. The value of *Whole 1* was between 11 and 15, *Whole 2* between 5 and 9, and the *Difference* was either 2 or 3.

We used a segmented self-presentation procedure displaying the text line by line on the screen when participants pressed the spacebar. Below, a question appeared: "Given the data provided, is it possible to find the solution?" followed by two possible choices: "(A) No, there is not enough information to find the solution." "(B) Yes, and the following solution is correct:" (followed by, in the case of the marble problem: "14 – 2 = 12. Lucy has 12 marbles in total"). A solution was proposed for each problem, and it was up to the participants to assess whether it was valid or whether the problem was unsolvable.

Table 2 Example of target problems used in the study. Changes introduced from Gros el al.'s (2017) problem statements are italicized in the table for the sake of clarity, but they were not made apparent in the experiment. Translated from French

Cardinal target problems	Ordinal target problems
Paul has *a certain amount of* red marbles He also has blue marbles In total, Paul has 14 marbles Jolene has as many blue marbles as Paul, and some green marbles She has two green marbles less than Paul has red marbles How many marbles does Jolene have?	Sofia travelled *for a certain time* Her trip started during the day Sofia arrived at 14 h Fred left at the same time as Sofia Fred's trip lasted 2 h less than Sofia's What time was it when Fred arrived?
In the store, Anthony wants to buy a ruler *costing a certain price* He also wants a notebook In total, that will cost him 14 dollars Julie wants to buy the same notebook as Anthony, and an eraser The eraser costs 2 dollars less than the ruler How much will Julie have to pay?	Slouchy Smurf is *a certain height* He climbs on a Smurf table He now attains the height of 14 cm Grouchy Smurf climbs on the same table as Slouchy Smurf Grouchy Smurf is 2 cm shorter than Slouchy Smurf What height does Grouchy Smurf attain when he climbs on the table?
Joe takes a Russian dictionary weighing *a certain weight* He also takes a Spanish dictionary In total, he is carrying 14 kg of books Lola takes Joe's Spanish dictionary and a German dictionary The German dictionary weighs 2 kg less than the Russian dictionary How many kilograms of books is Lola carrying?	Katherine took the elevator and went up *a certain number of floors* She left from the floor where the gym is She arrived to the 14th floor Yohan also took the elevator from the floor where the gym is He went up 2 floors less than Katherine What floor did Yohan arrive to?

Results

Data collected for both studies are available online (https://osf.io/fxgqh/?view_only=ed1374ef4d204c90a0cb03a30cb0a099). The dependent variable was the proportion of correct answers for solvable problems (see Fig. 2). Because multiple binary data points were recorded in a repeated design (each participant provided a binary answer to three ordinal and three cardinal solvable problems), the use of repeated measures ANOVA was deemed inappropriate and replaced by a mixed model (Hector, 2015). We used a generalized linear mixed model with a binary distribution, with the cardinal versus ordinal semantic nature of the problems as a fixed factor, and participants as a random effect. In line with our hypothesis, lay adults performed significantly better on ordinal (81.18%) than on cardinal problems (46.67%); $z = 7.84$, $p < .001$, $R^2_{GLMM(c)} = .29$.[2] Additionally, looking at individuals' response patterns showed us that 65.9% of the participants made fewer mistakes on ordinal than on cardinal problems, 11.8% made no mistakes at all, 15.3% made the same number of mistakes in cardinal and in ordinal problems, and only 7.1% made more mistakes on ordinal than on cardinal problems.

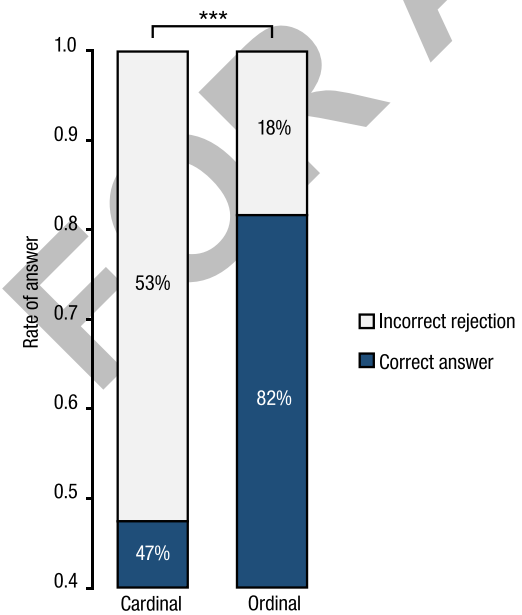

Fig. 2 Distribution of adults' answers. *** $p < .001$

Further analyses were conducted on participants' response times (RTs) on solvable problems that had been successfully identified as such by the participants (see Fig. 3). Because the number of correct answers could vary from 0 to 6 for each participant, the number of RT data points varied accordingly, and the use of repeated-measures ANOVA was again deemed inappropriate (Hector, 2015). A linear mixed model with subjects as a random effect and semantic nature of the problems as a fixed factor showed that participants took more time to correctly solve cardinal ($M = 34.05$, $SD = 18.78$) than ordinal problems ($M = 26.85$, $SD = 12.49$), χ^2 (1) = 29.14, $p < .001$, $R^2_{LMM(c)} = .44$. Additionally, we studied the participants' individual response patterns to identify whether different participant profiles existed. For each participant, we computed the difference between their mean RTs on correctly solved cardinal and ordinal problems (see Fig. 4) and we performed Hartigan's dip test for unimodality versus multimodality on the resulting distribution (Hartigan & Hartigan, 1985). The

[2] Conditional R^2 are reported in lieu of η^2 for the mixed models in this paper, since no satisfactory method is currently available to estimate effect sizes on mixed models (Westfall, Kenny, & Judd, 2014).

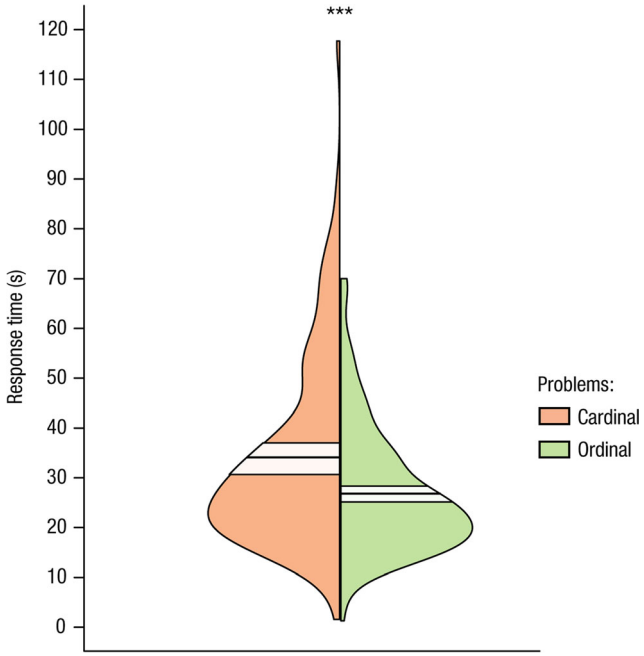

Fig. 3 Violin plot of adults' response times (RTs) on correctly identified solvable problems. Middle bars indicate mean RTs; upper and lower bars indicate margins of .95 confidence intervals. *** $p < .001$

analysis failed to reject the null hypothesis that participants' responses came from a unimodal distribution (D = .028, p = .94), thus providing no empirical ground to assume that the distribution of response times was multimodal.

Discussion

The difference in performance between cardinal and ordinal problems indicates that despite their expertise regarding basic subtractions, the adults' answers were significantly influenced by the semantic content of the problem statements. This confirms previous results obtained with the "complete" version of the problems that could be solved either with the three-step algorithm or with the one-step algorithm (Gamo et al., 2010; Gros et al., 2017). Here, we showed that the strategy imbalance observed in these previous studies was not an effect of mere preference for one strategy over another, but an actual impossibility to identify the relevance of the one-step algorithm on cardinal problems, as attested by the fact that on these problems more than half of the participants rejected a perfectly valid solution, despite only needing to check its validity. Regarding RTs, the fact that correct answers took more time on cardinal problems suggests that recognizing the solution to a problem evoking aspects of world semantics seemingly incompatible with the solution required an extra processing step. This is also supported by the fact that there was no significant difference in length between cardinal and ordinal problems. This is in line with the recoding process we predicted. These results show that the semantic content of a problem can prevent university-educated adults from recognizing a simple subtraction as the solution to a problem whose mathematical structure is undoubtedly within their level of expertise. We designed a second study to identify whether such effects would remain with expert mathematicians, known to be especially accustomed to abstract reasoning.

Study 2

Methods

Participants

We recruited 25 experts (two women, mean age = 23.59 years, *SD* = 2.81) who had successfully passed the entrance exam of the Science section at the École Normale Supérieure (ENS Ulm) in Paris. This exam is considered as the most demanding

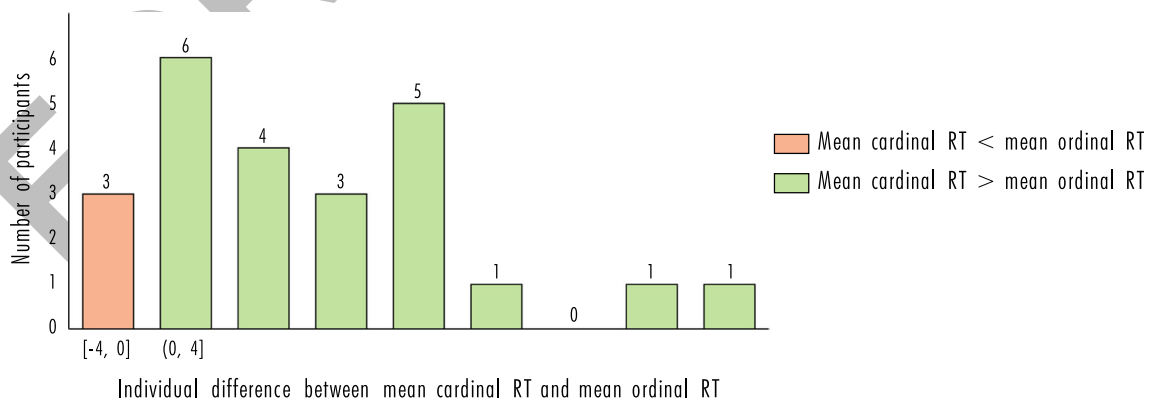

Fig. 4 Distribution of individual differences between cardinal response time (RT) and ordinal RT on correctly solved problems. Bins below the zero value indicate participants whose ordinal RT were higher than their cardinal RT on average, whereas bins above zero indicate participants whose ordinal RT were lower than their cardinal RT on average

one in France, with an entrance rate of 2.02% among university-educated participants ("SCEI Statistics", 2017). The ENS ranked second in Times Higher Education's World University Rankings 2016–2017 for Best Small University (Bhardwa, 2017). Although the population sample was smaller than in the first study due to the number of graduates from École Normale Supérieure being limited, sample size was deemed sufficient using uncertainty and publication bias correction on results from a previous study (Gros et al., 2016), following Anderson et al.'s recommendations (2017).

Materials and procedure

Materials and procedure were identical to that of Study 1.

Results

As in Study 1, we analyzed the proportion of correct answers on solvable problems (see Fig. 5) with a generalized linear mixed model. Experts had a higher success rate on ordinal (94.67%) than on cardinal problems (76.00%); $z = 2.99$, $p = .0028$, $R^2_{GLMM(c)} = .25$. Additionally, a comparison with Study 1 showed that Study 2 experts' performance (85.33%) was significantly higher than Study 1 adults' performance (63.92%), which was another confirmation of their outstanding expertise in mathematics; $z = 4.49$, $p < .001$, $R^2_{GLMM(c)} = .33$. Looking at individuals' response patterns also indicated that 52.0% of the participants made fewer mistakes on ordinal than on cardinal problems, 36.0% made no mistakes at all, 4.0% (one participant) made the same number of mistakes in cardinal and in ordinal problems, and only 8.0% made more mistakes on ordinal than on cardinal problems.

Analyses were conducted on participants' RTs for correctly identified solvable problems (see Fig. 6). As in Study 1, we used a linear mixed model that showed that experts took significantly more time to correctly solve cardinal problems ($M = 26.58$, $SD = 14.03$) than ordinal problems ($M = 19.45$, $SD = 8.18$), as predicted by our world semantics hypothesis; $\chi^2(1) = 18.65$, $p < .001$, $R^2_{LMM(c)} = .37$. Unsurprisingly, experts' RTs on correct answers were significantly shorter ($M = 22.63$, $SD = 11.68$) than in Study 1 ($M = 29.50$, $SD = 15.48$); $\chi^2(1) = 7.68$, $p = .0056$, $R^2_{LMM(c)} = .46$. As in Study 1, the computation of individual differences in RTs between cardinal and ordinal problems showed no sign of multimodality (see Fig. 7), and Hartigan's dip test for unimodality versus multimodality failed to reject the null hypothesis of unimodality ($D = .048$, $p = .96$).

Discussion

Despite their superior performances, high-level mathematicians were still significantly influenced by world semantics. Their performance dropped significantly on cardinal problems, and correct answers required more time on average on cardinal than on ordinal problems. Therefore, despite their proficiency in abstract mathematical reasoning, expert mathematicians failed to disregard irrelevant non-mathematical information when solving the problems, as hypothesized.

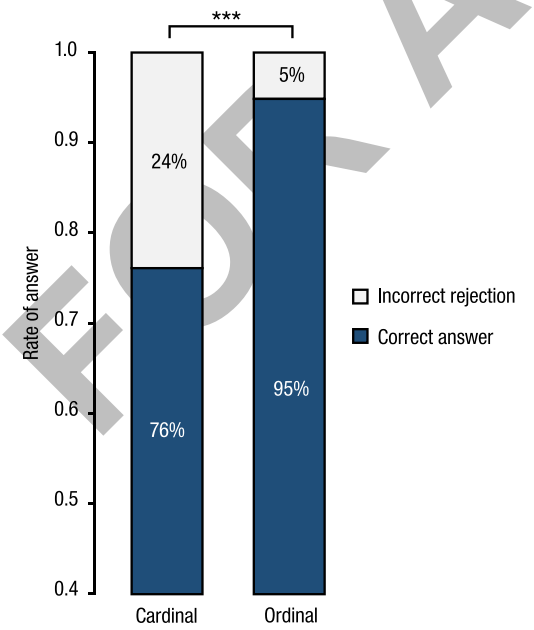

Fig. 5 Distribution of experts' answers. ** $p < .01$

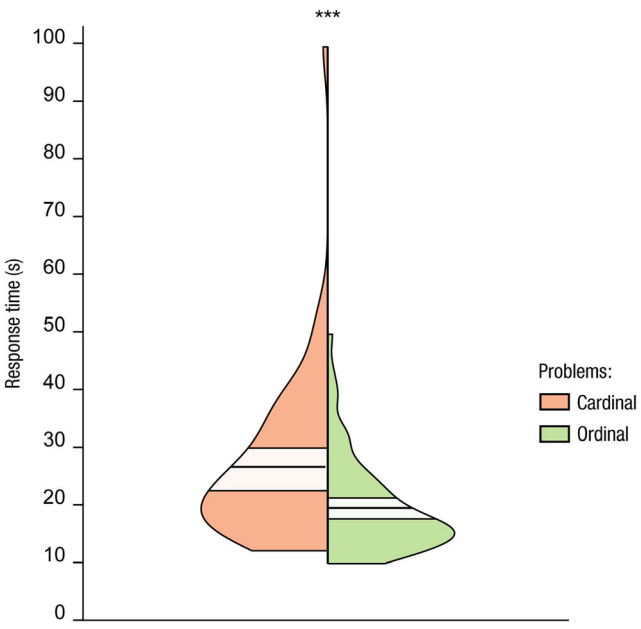

Fig. 6 Violin plot of experts' response times (RTs) on correctly identified solvable problems. Middle bars indicate mean RTs; upper and lower bars indicate margins of .95 confidence intervals. *** $p < .001$

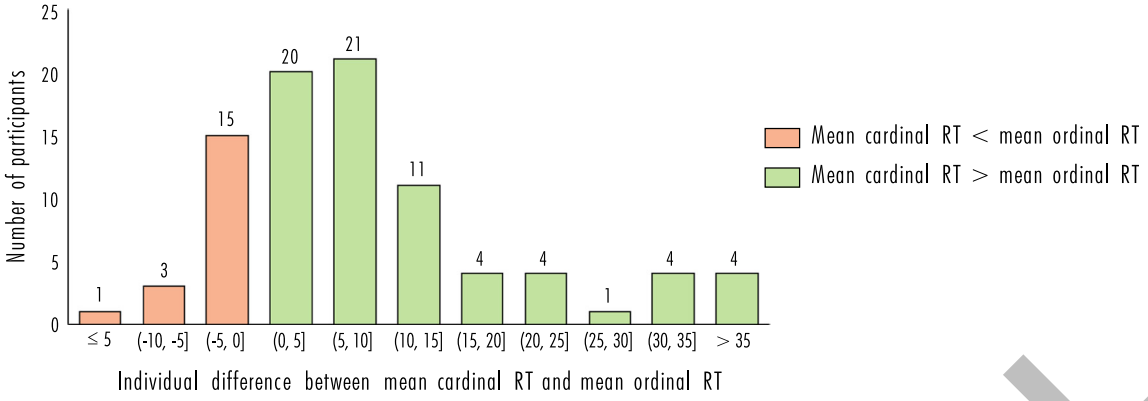

Fig. 7 Distribution of individual differences between cardinal response time (RT) and ordinal RT on correctly solved problems. Bins below the zero value indicate participants whose ordinal RT were higher than their

General discussion

In this paper, we sought to demonstrate that irrelevant aspects of our non-mathematical knowledge evoked by the semantic content of a problem statement can lead both adults and mathematics experts to encode the problem in such a way that they would erroneously consider valid solutions as incorrect. Indeed, participants failed to identify the solvability of subtraction problems admitting a single-step solution significantly more often when the world semantics they evoked conflicted with the relevant mathematical information, than when the two were congruent. Additionally, correct answers took more time in the conflicting than in the congruent case for both populations, suggesting that the initial spontaneous representation triggered by the semantic content of the problem statement had to be recoded. Although they achieved higher performances overall, high-level experts still rejected several perfectly valid solutions: they fell prey to robust effects of world semantics that current theories of expertise do not account for.

There is a much larger body of literature describing in what terms experts excel in their field than there are studies revealing experts' shortcomings. However, as Chi (2006, p. 23) stressed, "it is equally important to understand how experts fail," which was one of the goals of this paper. A few limitations have already been shown to occasionally affect experts' excellence (see Chi, 2006, for a review). For instance, experts' proficiencies are limited to their domain of expertise (Ericsson & Lehmann, 1996) and they lack adaptability to irregular situations whose structures differ from what they expect (Sternberg & Frensch, 1992). They have even been shown to gloss over details (Voss, Vesonder, & Spilich, 1980), which paradoxically suggests that they should be good at ignoring surface properties unrelated to the formal structure of the problems. More recent works have even hinted at biases slowing down experts within their own domain of expertise (Goldberg & Thompson-Schill, 2009; Obersteiner, Van Dooren, Van Hoof, & Verschaffel, 2013). However, we believe none of these accounts would have predicted our results, since they do not explain how mathematically irrelevant contextual information may significantly hinder experts' abstract reasoning on problems within their very field of expertise, to the extent that they would not identify the validity of the solution handed out to them. Here, mathematical experts failed to do what they are good at: engaging in abstract reasoning on concrete entities to find a single-step solution. Our results suggest that when mathematical knowledge and world semantics conflict with one another, masters of abstraction can run into a concrete wall.

This effect is understandable since world semantics and mathematical knowledge often (although not always) naturally align with each other, which explains how some superficial cues are highly correlated with deeper principles (Bassok, Pedigo, & Oskarsson, 2008; Blessing & Ross, 1996). It follows that solvers rely on those cues at all levels and tend to make mistakes when world and mathematical semantics do not align. Overall, it seems that these effects of semantic (in)congruence between world semantics and mathematical knowledge have been greatly undermined on the account of mathematics being an inherently abstract domain in which rules and concepts are valid independently from the objects they are applied to. Our results show how prevalent the influence of world knowledge is on arithmetic reasoning, even among the individuals who should be the least subject to it. This suggests that experts will never be completely freed from the influence of world knowledge; having an outstanding level in mathematics is not enough to systematically perceive that 14 – 2 = 12.

Acknowledgments We sincerely thank Pernille Hemmer and two anonymous reviewers whose insightful feedback helped improve and clarify this manuscript. We also acknowledge gratefully Pierre Barrouillet, Katarina Gvozdic, and Maxime Maheu for helpful comments on previous versions of this work.

This research was supported by grants from the Regional Council of Burgundy, Pari Feder Grants (20159201AAO050S02982 & 20169201AAO050S01845, JPT), from the Experimental Fund for the Youth and French Ministry of Education (HAP10-CRE-EXPE-S1, ES), and from the French Ministry of Education and Future Investment Plan (CS-032-15-836-ARITHM-0, ES). HG was further supported by a doctoral fellowship from the Paris Descartes University.

Open practices statement The data and materials for all experiments are available at (https://osf.io/fxgqh/?view_only=ed1374ef4d204c90a0cb03a30cb0a099).

References

"SCEI Statistics" (2017) Retrieved from http://www.scei-concours.fr/statistiques.php.

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562.

Bassok, M. (2001). Semantic alignments in mathematical word problems. In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 401–433). Cambridge, MA: MIT Press.

Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology*, *35*(2), 99–134.

Bassok, M., Pedigo, S. F., & Oskarsson, A. T. (2008). Priming addition facts with semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 343–352.

Bassok, M., Wu, L. L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory and Cognition, 23*, 354–367.

Bhardwa, S. (2017). International Student Table 2017: Top 200 Universities. *Times Higher Education*.

Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 792–810.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15,* 179–202.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R.S. Siegler (Ed.) *Children's thinking: What develops?*, (pp. 73–96). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 121–130). Cambridge: Cambridge University Press.

Davidson, J. E., & Sternberg, R. J. (Eds.). (2003). *The psychology of problem solving*. New York, NY: Cambridge University Press.

Davis, P., Hersh, R., & Marchisotto, E. A. (2011). *The mathematical experience, Study edition*. Boston, MA: Birkhäuser.

De Groot, A. D. (1965). *Thought and choice in chess*. The Hague, Netherlands: Mouton.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. New York, NY: Oxford University Press.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*(1), 273–305.

Gamo, S., Sander, E., & Richard, J. F. (2010). Transfer of strategy use by semantic recoding in arithmetic problem solving. *Learning and Instruction*, *20*(5), 400–410.

Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science*, *20*(4), 480–487.

Gros, H., Sander, E., & Thibaut, J. P. (2016). "This problem has no solution": When closing one of two doors results in failure to access any. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1271–1276). Austin, TX: Cognitive Science Society.

Gros, H., Thibaut, J. P., & Sander, E. (2015). Robustness of semantic encoding effects in a transfer task for multiple-strategy arithmetic problems. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 818–823). Austin, TX: Cognitive Science Society.

Gros, H., Thibaut, J. P., & Sander, E. (2017). The nature of quantities influences the representation of arithmetic problems: Evidence from drawings and solving procedures in children and adults. In R. Granger, U. Hahn, & R. Sutton (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp 439–444). Austin, TX: Cognitive Science Society.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84.

Hector, A. (2015). *The new statistics with R: An introduction for biologists*. Oxford, UK: Oxford University Press.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser, M. J. Farr (Eds.), *The nature of expertise* (pp. 311–342). Hillsdale, NJ: Erlbaum

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.

Obersteiner, A., Van Dooren, W., Van Hoof, J., & Verschaffel, L. (2013). The natural number bias and magnitude representation in fraction comparison by expert mathematicians. *Learning and Instruction*, *28*, 64–72.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 629–639.

Russell, B. (1903). *Principles of mathematics*. Cambridge, UK: Cambridge University Press.

Sternberg, R. J., & Frensch, P. A. (1992). On being an expert: A cost-benefit analysis. In *The psychology of expertise* (pp. 191–203). New York, NY: Springer.

Thevenot, C., & Barrouillet, P. (2015). Arithmetic word problem solving and mental representations. In R. Cohen Kadosh, & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 158–179). Oxford, UK: Oxford University Press.

Verschaffel, L., De Corte, E., & Vierstraete, H. (1999). Upper elementary school pupils' difficulties in modeling and solving nonstandard additive word problems involving ordinal numbers. *Journal for Research in Mathematics Education*, *30*(3), 265–285.

Vicente, S., Orrantia, J., & Verschaffel, L. (2007). Influence of situational and conceptual rewording on word problem solving. *British Journal of Educational Psychology*, *77*(4), 829–848.

Voss, J. F., Greene, T. R., Post, T. A., & Penner, B. C. (1983). Problem-solving skill in the social sciences. In *Psychology of learning and motivation* (Vol. 17, pp. 165–213). New York, NY: Academic Press.

Voss, J. F., Vesonder, G. T., & Spilich, G. J. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior*, *19*(6), 651–667.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*, 2020–2045.