Are content effects out of sight? An eye-tracking study of arithmetic problem solving

Hippolyte Gros (hippolyte.gros@unige.ch),
Emmanuel Sander (emmanuel.sander@unige.ch),
& Jean-Pierre Thibaut (jean-pierre.thibaut@u-bourgogne.fr)

Abstract

Evidence suggests that general, non-mathematical knowledge about the entities described in an arithmetic word problem may interfere with its encoding. We used behavioral and eye-tracking measures to investigate how the use of specific quantities may foster a cardinal representation of the numbers mentioned in a problem, whereas other quantities may favor an ordinal representation instead. We asked 50 pre-service teachers to complete a solution validity assessment task. We compared participants' gaze patterns on isomorphic problems to gather insights into their encoded representations. On problems featuring cardinal quantities, we found that specific sentences describing elements relevant in a cardinal understanding of the problems but irrelevant otherwise were looked at longer and were the focus of a higher number of backward eye movements. Additionally, an increase in pupil dilation on correctly solved cardinal problems supported the idea that participants need to engage in a recoding process when facing semantic incongruence.

Keywords: arithmetic word problems; encoding effects; eye tracking; mathematical cognition; problem solving

Introduction

Mathematical word problems are infamously difficult, and many students struggle with the delicate exercise consisting in applying abstract mathematical notions to concrete, daily-life situations (Daroczy, Wolska, Meurers, & Nuerk, 2015; Verschaffel, Greer, & De Corte, 2000). But what makes some mathematical word problems so hard to solve? Several lines of work have looked at the interaction between linguistic and numerical factors to account for the interpretative processes at play in mathematical word problem solving (Thevenot & Barrouillet, 2015).

Notably, the issue of the underlying representations accounting for the strategies developed by students to solve the problems they encounter has been a recurring question in the literature. It has for example been proposed that students use problem schemata (Kintsch & Greeno, 1985; Schank & Abelson, 1977) or mental models (Johnson-Laird, 1983; Staub & Reusser, 1995). It has also been suggested that general semantic knowledge about the entities featured in a problem interfere with its solving process, by means of an interpreted structure describing one's interpretation of the situation depicted in the problem (Bassok, 2001). More recently, the SECO framework (Gros, Thibaut, & Sander, 2020a), suggested that an initial semantic representation is encoded based on the problem statement and on the solver's general, non-mathematical knowledge about the entities it features. This approach notably predicts that an inappropriate encoding of a given problem statement may sometimes be semantically recoded into a new representation, in an attempt to overcome a dead end and find the solution to an arduous problem. Following SECO's predictions, our paper investigates the role of prior knowledge on the encoding, recoding, and solving of arithmetic word problems by studying the perception of cardinality and ordinality among pre-service teachers, using behavioral and eye-tracking data.

Cardinal encoding versus ordinal encoding

In common usage, ordinal numbers describe the numerical position of an object in an ordered sequence (i.e. 1st, 2nd, 3rd, etc.), whereas cardinal numbers refer to the general concept of quantity by designating the total number of entities within a set (Wasner, Moeller, Fischer, & Nuerk, 2015). The difference between these two meanings of numbers is central to the notion of number itself (Fuson, 1988) and the understanding of cardinality and ordinality has been the focus of numerous studies (e.g. Colomé & Noël, 2012; Lyons, Vogel, & Ansari, 2016). However, until recently, the importance that this distinction holds for the representation of word problems had received scant attention in the literature. A line of work has aimed to fill this gap, by targeting the cardinal and ordinal representations of arithmetic word problems. Gamo, Sander, and Richard (2010) found that students' choice of solving algorithms varied between number-of-element problems, price problems and age problems. They suggested that while number-of-element problems and price problems feature unordered elements that tend to be represented as sets and subsets, age problems are more easily represented along an axis (a timeline) and the apparent order between the age values facilitates the use of a different solving strategy. This distinction was framed in terms of *ordinal* and *cardinal* encodings, thus introducing the idea that quantities emphasizing the cardinal aspect of numbers led to different representations than quantities emphasizing their ordinal aspect.

To investigate this distinction in a systematic way, new arithmetic word problems were created using different types of quantities (Gros, Thibaut, & Sander, 2017; Gros, Thibaut, & Sander, 2020b). Figure 1 provides a graphical summary of the expected encoding of these problems. The problems all shared the same abstract mathematical structure (Figure 1, box 1.), but they were implemented either with cardinal quantities (Figure 1, box 2.a) or with ordinal quantities (Figure 1, box 3.a). Consider for instance the cardinal problem reproduced in Figure 1, box 2.b: by mentioning collections of unordered marbles, this problem is expected to emphasize the cardinal aspect of numbers, and thus to elicit a cardinal encoding of the situation (Figure 1, box 2.c). This representation fosters the idea that to find the number of

marbles that Jolene has (*Whole 2*), one needs to add up the number of blue marbles she has (*Part 2*) and the number of green marbles she has (*Part 3*). This representation is thus semantically congruent with a 3-step algorithm (Figure 1, box 3.c) consisting in calculating the value of *Part 2* (*Whole 1 – Part 1 = Part 2*), and adding it to the value of *Part 3* (*Part 1 – Difference = Part 3*), to find the solution to the problem (*Part 2 + Part 3 = Whole 2*).

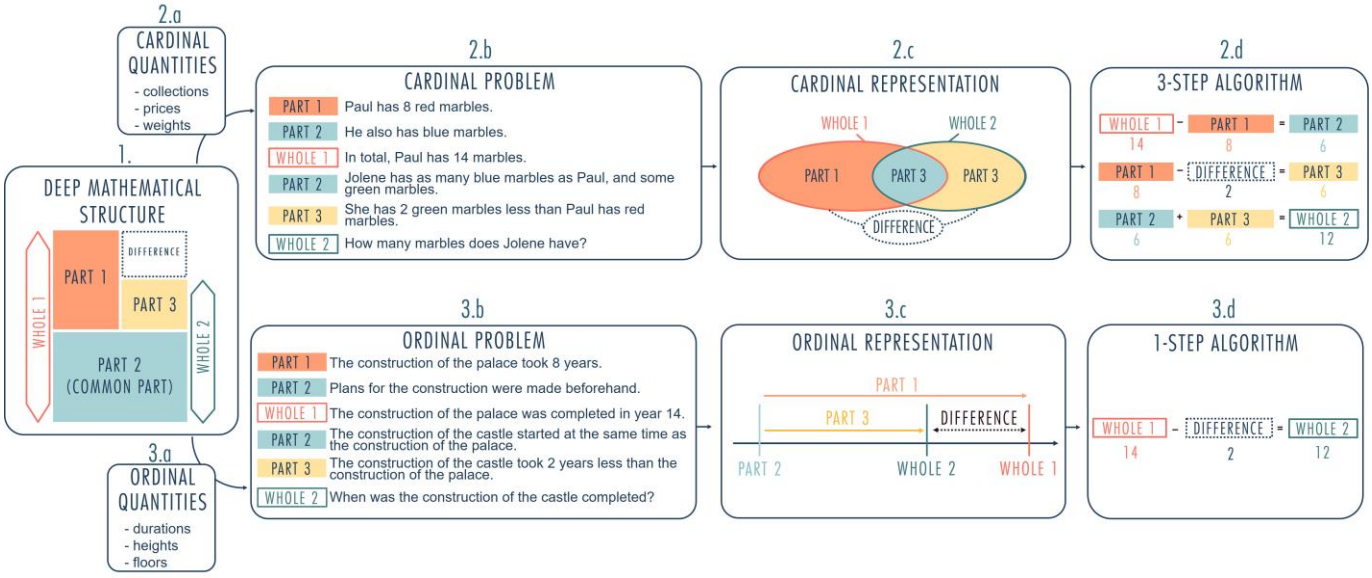

Figure 1: Implementation of the mathematical structure with ordinal versus cardinal quantities, leading to different problem statements, representations, and strategy use.

On the other hand, the duration problem in box 3.b describes a situation that can easily be represented along an axis (a timeline), and it is thus thought to evoke an ordinal encoding (Figure 1, box 3.c). This representation facilitates the understanding that since the construction of the palace and that of the castle started at the same time, and since the construction of the castle took 2 years less than the construction of the palace, then the castle was completed 2 years earlier than the palace. Thus, this inference makes it easier to use a 1-step algorithm to find the *Whole 2* value: *Whole 1 – Difference = Whole 2* (see Figure 1, box 3.d).

Depending on the cardinal versus ordinal nature of the quantities used, participants were thought to construct a different encoding of the situation, which led them to one of the two solving algorithms. Both drawing productions elicited by Gros et al. (2017), and participants' report of the algorithms they used supported this claim. To evaluate the robustness of these effects, Gros, Sander, and Thibaut (2019) designed a modified version of these problems, in which the value of *Part 1* was not provided. For instance, the sentence "Paul has 8 red marbles" was replaced by "Paul has some red marbles", and the sentence "The construction of the palace took 8 years" was replaced by "The construction of the palace took a certain time". It thus became impossible to use the 3-step algorithm (Figure 1, box 2.d) since that required knowing the value of *Part 1*, and the only algorithm left to solve the problems was the 1-step algorithm (Figure 1, box 3.d). Gros et al. created a solution-assessment task, in which these problems were presented accompanied by their solution, and participants had to decide whether the solution was correct or whether the problems could not be solved. They presented this task to lay adults and to expert mathematicians and found, as predicted, that in both cases their expertise was not enough to prevent the influence of the cardinal versus ordinal distinction: participants made more errors and took longer to solve cardinal problems. In the current study, we intend to use an eye-tracking setup to get a finer understanding of the difference between these representations.

Eye tracking as an index of reasoning processes

From an educational standpoint, there has been an increasing amount of literature using eye tracking to better understand students' learning processes (Lai et al., 2013). In the study of mathematical reasoning, eye tracking has been used to pinpoint the integration of relevant information while performing calculations or solving math problems (Curtis, Huebner, & LeFevre, 2016; Merkley & Ansari, 2010). However, a surprisingly low number of studies have resorted to this methodology to understand mathematical *word* problem solving (Strohmaier, Tatsidou, & Reiss, 2018).

In fact, ever since De Corte and Verschaffel's (1986) seminal work on the matter, we are aware of less than a dozen studies who looked at mathematical word problem solving using eye movement recording. For instance, De Corte, Verschaffel and Pauwels (1990) used eye tracking to discriminate between the initial read-through of arithmetic word problems and the subsequent time spent rereading the problem statement. Then, Verschaffel, De Corte, and Pauwels (1992) showed that students' longer response times on problems featuring relational terms inconsistent with their solving algorithms were due to a longer time spent on the initial reading of the problems' first sentences. Similarly, eye-tracking has been used to compare high-performing and

low-performing students' reading patterns (Hegarty, Mayer, & Green, 1992; Hegarty, Mayer, & Monk, 1995). Later, van der Schoot, Bakker Arkema, Horsley and van Lieshout (2009) focused on regressive eye movements to evaluate how the strategies of successful and less successful problem solvers differed. Finally, Dewolf et al. (2015) used looking time analysis to investigate how often students looked at representational illustrations accompanying word problems.

Thus, to the best of our knowledge, most of the research conducted on arithmetic word problems using eye-tracking methodology has focused on looking times, with some works counting backward eye-movements to identify specific strategies. In our study, we intend to use both metrics to get a finer understanding of the differences between cardinal and ordinal problems, as well as a third one selected to evaluate participants' effort in the task: pupil dilation.

Pupillometry concerns the measure of pupil dilation over time. Its use in research was initiated by Hess and Polt (1964), who found that pupils tended to dilate when individuals were asked to solve multiplication non-word problems of increasing difficulty. Subsequent works discovered that pupil diameter increased with memory load (Goldinger & Papesh, 2012) and with task demand in general (Beatty, 1982), which makes it a valuable index to evaluate participants' effort variations when solving arithmetic word problems.

Current study

While it seems that the difference between cardinal and ordinal problems runs deep enough to interfere even with math experts' understanding of arithmetic word problems (Gros et al., 2019), the question remains as to what exactly this distinction entails in terms of solving processes. In this study, we strove to probe participants' representations using direct measures that would not solely rely on verbal or written productions.

We analyzed the gaze patterns and pupil dilation of 50 participants engaging in the solving of problems similar to those used in Gros et al. (2019), where the value of *Part 1* was not provided to the participants, and we made the following predictions. First, the total looking time (visit duration) spent on each line of the problems should vary between cardinal and ordinal problems. Since a cardinal encoding is supposed to foster the calculation of *Part 2* and *Part 3* to find *Whole 2*, we expected that cardinal problems would lead to longer visit durations on the lines referring to *Part 2* and *Part 3*, compared to ordinal problems. Second, since the values of *Part 2* and *Part 3* are not provided in the problem statements but are nevertheless deemed necessary by participants who construct a cardinal encoding, then backward eye movements to the lines referring to these two quantities should be more frequent on cardinal problems. Third, since participants who manage to solve cardinal problems are thought to engage in a costly semantic recoding process (Gros et al., 2019), then correctly solving a cardinal problem should result in an increase in pupil diameter whereas solving an ordinal problem should not. In addition, this study aimed at replicating two results from Gros et al. (2019): cardinal problems should be solved less frequently and require a longer response time than ordinal problems.

Methods

Participants. Participants were 50 pre-service teachers (41 women, M = 27.22 years, SD = 13.95) recruited from the Educational Sciences program at the University of Geneva. All of them spoke French fluently and volunteered in exchange for course credit.

Materials. The arithmetic word problems used in this experiment were taken from the 12 problems created in Gros et al. (2019), to which 6 new problems were added, constituting a pool of 18 problems to choose from. All problems were written in French. Each participant was presented with a random selection of 12 target solvable problems: 6 with cardinal quantities (2 collection problems, 2 price problems, and 2 weight problems) and 6 with ordinal quantities (2 duration problems, 2 height problems, and 2 floor problems). A within-subject design was used to allow for within-subject comparisons between performance on cardinal and ordinal problems. In addition, we introduced 6 unsolvable filler problems that were similar to the target problems but did not provide any value for *Whole 2*, which meant that they could not be solved with any algorithm. Order of target and filler problems was randomized between participants. The numerical values used were randomized across problems.

Procedure. The stimuli were presented on a 23.8" monitor. Participants were seated approximately 65 centimeters from the monitor in a soundproofed experimental room. The eye movements were registered with a *Tobii Pro Spectrum* eye tracker. There was no window to avoid any natural light fluctuation. The first screen displayed the instructions for the experiment. They were provided the following instructions: "You will be presented with a series of arithmetic problems. 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. Press the space bar when you are ready to start". A fixation cross was displayed for 3 seconds before each problem.

Each problem screen comprised 6 lines of text composing the problem statement, and a separate insert displaying the response choices. The text was written in size 18, with a line spacing of 3.7 to ensure that minor inaccuracies of the eye gaze estimation would not be detrimental. The response insert presented two possible choices. Choice "A" was the solution to the problem (e.g. "14 – 2 = 12. Jolene has 12 marbles."). Choice "B" stated: "There is not enough information to find the solution". Participants answered each problem using two keys on a keyboard placed in front of them. A typical session lasted between 20 and 30 minutes.

Results

Success Rates. We looked at participants' failures and successes on solvable problems. Since each participant gave a binary answer to 6 cardinal and 6 ordinal problems, we used a generalized linear mixed model (GLMM) with a binary distribution to account for the repeated measures in the experimental design. We used the cardinal versus ordinal nature of the problems as a fixed factor and participants as a random effect. The overall model successfully converged and had a total explanatory power of 12.60% (conditional R^2). In line with previous results, participants performed significantly worse on cardinal (51.51%) than on ordinal problems (82.4%); $z = 8.22$, $p < .001$.

Response Times. We looked at the RTs of correctly solved problems (see Figure 2). We had predicted that solving a cardinal problem would require a higher RT, due to an extra recoding step being necessary to find the solution. We used Tukey's method to remove 16 outliers ranged above and below 1.5 interquartile range. We analyzed participants' RTs using a linear mixed model with the cardinal versus ordinal nature of the problems as a fixed factor and participants as a random effect. The model successfully converged and explained 30.43% of the variance (conditional R^2). Within this model, an ANOVA using Satterthwaite's method for estimation of degrees of freedom revealed that there was a significant effect of the cardinal versus ordinal nature of the problems on the RTs of correctly solved problems ($F(1) = 25.24$, $p < .001$). Which indicates that solving a cardinal problem required more time on average ($M = 25.29$, $SD = 8.72$) than solving an ordinal problem ($M = 21.73$, $SD = 7.73$). The results from Gros et al. (2019) were thus replicated, both in terms of success rate and response times.

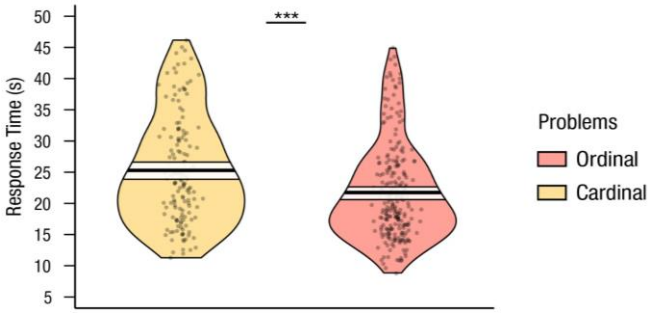

Figure 2: Pirate plot of RTs on cardinal and ordinal problems. Middle lines indicate mean RT, upper and lower lines indicate 95% confidence interval. *** $p < .001$.

Scoring of eye-fixation data. The sequence of eye fixations for each participant was recorded with the software *Tobii Pro Lab*. We partitioned the screen into 7 different areas of interest (AOIs): one for each problem line, and one dedicated to the response insert. The seven AOIs of equal height and width partitioned the entire screen.

Visit durations. We had predicted that, since cardinal problems are supposed to lead to a cardinal encoding, participants will spontaneously try to calculate the intermediate values of *Part 2* and *Part 3* to find the value of *Whole 2* (See Figure 1, box 2.d). Thus, participants should spend more time in sentences referring to these two subsets on cardinal problems, that is the lines referring to *Part 2* (lines 2 and 4) and *Part 3* (line 5; see Figure 1, box 2.b).

We extracted the total visit duration per AOI for each participant. Since each participant's gaze was recorded on 12 different problems, we analyzed the visit duration using a GLMM with visit duration as the dependent factor, participants as a random effect, the line number as a fixed effect and the cardinal versus ordinal nature of the problems as a fixed effect. The model successfully converged and had a total explanatory power of 31.92% (conditional R^2). Within this model, an ANOVA using Satterthwaite's approximation for the degrees of freedom revealed that the cardinal versus ordinal nature of the problems had a significant effect ($F(1) = 53.74$, $p < .001$), as well as the line number ($F(6) = 202.69$ $p < .001$). The interaction between those two fixed effects was significant as well ($F(6) = 12.29$, $p < .001$). In accordance with our hypothesis, we computed orthogonal contrasts using least square means to identify whether participants did visit the lines referring to *Part 2* and *Part 3* longer on cardinal than on ordinal problems: lines 2, 4 and 5.

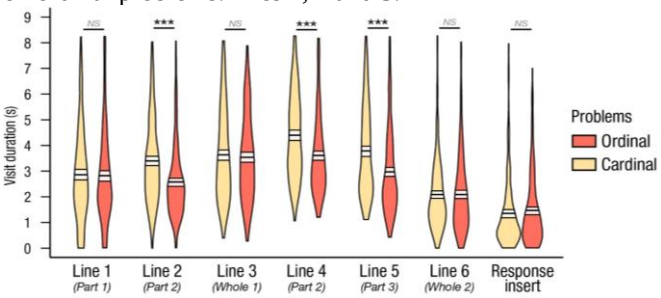

Figure 3: Visit duration per problem line

Results revealed that participants spent a longer time visiting line 2 on cardinal problems ($M = 3.40$ seconds, $SD = 1.64$) than they did on ordinal problems ($M = 2.58$ seconds, $SD = 1.43$); $t(3922) = 6.49$, $p < .001$. They also spent longer time on line 4 on cardinal problems ($M = 4.40$ seconds, $SD = 1.66$) than on ordinal problems ($M = 3.61$ seconds, $SD = 1.60$); $t(3924) = 6.13$, $p < .001$. Finally, they spent a longer time on the 5th line of cardinal problems ($M = 3.79$ seconds, $SD = 1.75$) than on that of the ordinal problems ($M = 2.97$ seconds, $SD = 1.67$); $t(3923) = 6.56$, $p < .001$. On the other hand, there was no significant visit duration difference between cardinal and ordinal problems on lines 1, 3, 6 nor on the response insert (see Figure 3) [1]; $0.02 \leq$ t-value ≤ 0.96, $.33 \leq p \leq .98$.

[1] Due to two thirds of the problems coming from a previous study (Gros et al., 2019), we could not perfectly control for the word length of every line. While there was no length difference in lines 2 to 6, there was a higher number of words in line 1 of ordinal problems ($M = 6.89$, $SD = 2.15$) as compared to line 1 of cardinal problems ($M = 10.22$, $SD = 2.82$); $t(16) = 2.82$, $p < .05$. This difference was not deemed problematic since our hypotheses focused on lines 2, 4 and 5, and since there was no significant difference of the visit duration on line 1 between cardinal and ordinal problems.

Regressions. We investigated participants' number of backward eye movements (regressions). Since each problem line presented a new piece of information, we could infer which pieces of information participants were going back to when trying to solve the problems. For each trial, we calculated the total number of backward eye movements to each line. We had predicted that participants would make more regressions to the lines mentioning *Part 2* (lines 2 and 4) and *Part 3* (line 5) in their search for the missing values needed to use the 3-step algorithm.

We used a GLMM with number of regressions as the dependent factor, cardinal versus ordinal nature of the problems as a fixed factor, line number as a fixed factor and participants as a random effect. The model successfully converged with a total explanatory power of 24.75% (R^2_{cond}). Within this model, an ANOVA using Satterthwaite's estimation revealed that the effect of the cardinal versus ordinal nature of the problems was statistically significant ($F(1) = 140.11$, $p < .001$). There was also a main effect of the line number ($F(5) = 94.43$, $p < .001$). The interaction between these two fixed factors was significant ($F(5) = 16.36$, $p < .001$). In accordance with our hypothesis, we computed orthogonal contrasts using least square means to identify whether participants did make more regressions to lines 2, 3 and 5 on cardinal problems than they did on ordinal problems (see Figure 4).

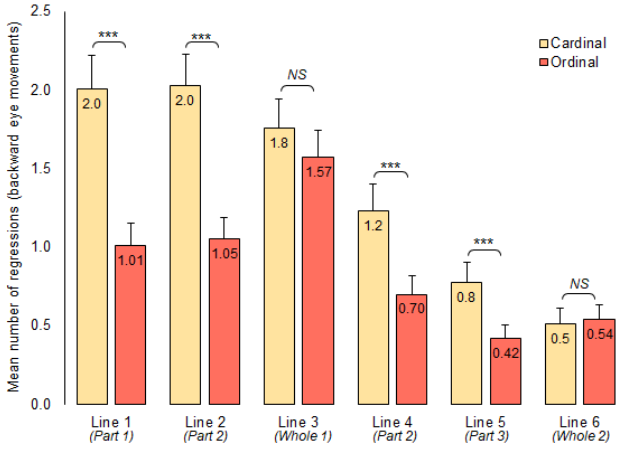

Figure 4: Mean number of regressions to specific lines. Error bars indicate 95% confidence intervals.

Results revealed that, as predicted, participants made a higher number of regressions to line 2 on cardinal problems (M = 2.03, SD = 1.73) than on ordinal problems (M = 1.05, SD = 1.16); $t(3306) = 9.39$, $p < .001$. Similarly, they made more regressions to line 4 on cardinal problems (M = 1.23, SD = 1.43) than on ordinal problems (M = 0.70, SD = 1.01); $t(3306) = 5.12$, $p < .001$. Finally, the number of regressions to line 5 was higher on cardinal problems (M = 0.77, SD = 1.11) than on ordinal problems (M = 0.42, SD = 0.76); $t(3306) = 3.43$, $p < .001$. The contrast analysis also revealed a difference that we had not anticipated: participants made a higher number of regressions to line 1 on cardinal problems (M = 2.00, SD = 1.83) than on ordinal problems (M = 1.01, SD = 1.24); $t(3306) = 9.63$, $p < .001$. There was no such difference between cardinal and ordinal problems on line 3 ($t(3306) = 1.75$, $p = .08$) nor on line 6 ($t(3306) = 0.31$, $p = .76$).

Pupillary dilatation. To evaluate the validity of the claim that participants need to semantically recode their initial representation of cardinal problems to find the solution, we looked at participants' pupil dilation in relation with their successes and failures in solving the problems. We measured participants' pupil diameter at each time step during each problem and contrasted it with their answers to the problems. We analyzed inter-trial change in pupil diameter using a GLMM with pupil diameter during fixations as the dependent variable. We used participants as a random effect, the cardinal versus ordinal nature of the problems as a fixed factor and the participants' response to the problems as a fixed factor. The model successfully converged and explained 85.21% of the variance (R^2_{cond}). Within this model, an ANOVA using Satterthwaite's approximation revealed that the cardinal versus ordinal nature of the problems had a significant effect ($F(1) = 68.87$, $p < .001$), indicating that pupil dilation differed between cardinal and ordinal problems. There was no main effect of the response provided by the participants ($F(1) = 0.38$, $p = .54$). There was however an interaction between the type of problem (cardinal/ordinal) and the response given by the participants (true/false): $F(1) = 5.73$, $p < .05$.

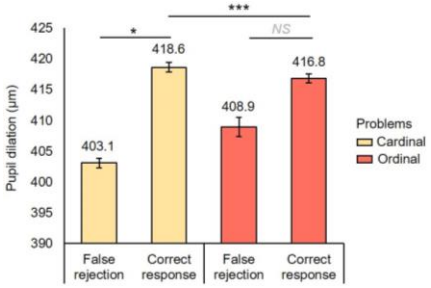

Figure 5: Pupil dilation on solvable problems. Error bars indicate upper margins of 95% confidence intervals.

In accordance with our hypothesis, we computed contrasts using least square means to identify whether participants' response was linked to their pupil dilation on cardinal and ordinal problems (see Figure 5). Correctly solving a cardinal problem was associated with a larger pupil diameter on average as compared to correctly solving an ordinal problem ($t(58838) = 5.34$, $p < .001$), which suggests that finding the solution to cardinal problems was more cognitively taxing than finding the solution to ordinal problems. Besides, a comparison of successes and failures revealed that participants' pupil diameter was significantly larger on correctly solved cardinal problems (M = 418.63 µm, SD = 53.14) than on failed cardinal problems (M = 403.08 µm, SD = 49.05); $t(58841) = 2.48$, $p < .05$. On the other hand, there was no such difference between correctly solved ordinal problems (M = 416.83 µm, SD = 54.18) and incorrectly rejected ordinal problems (M = 408.90 µm, SD = 55.88); $t(58841) = 1.09$, $p = .28$. This suggests an increase in cognitive load on cardinal problems correctly solved, but not on ordinal problems.

Discussion

In this paper, we gathered converging evidence from five different sources of information regarding what precisely happens when one's non-mathematical knowledge interferes with one's mathematical expertise in the encoding, recoding, and solving of arithmetic word problems. First, the success rate analysis confirmed previous results regarding the increased difficulty to perceive the validity of the 1-step algorithm on cardinal problems as compared to ordinal problems. Then, the difference in response times between correctly solved cardinal and ordinal problems was also replicated, supporting the hypothesis that one needs to engage in a semantic recoding step to construct a new representation compatible with the 1-step algorithm.

Third, by studying the visit duration on each line of the problem, we were able to take a closer look at what differentiates the encoding of cardinal and ordinal problems. We hypothesized that problems using cardinal quantities would lead participants to abstract a cardinal encoding of the situation emphasizing the set/subset structure of the situation depicted. Thus, in their attempts to find the value of *Whole 2*, we predicted that participants' first reaction would be to try to find the values of each of its subsets, that is, *Part 2* and *Part 3* (see Fig. 1). Our looking time analysis revealed that it was indeed the case, since lines 2, 4 and 5 were visited for a longer time on cardinal problems than on ordinal problems. Despite the lines presenting the same information in the same order across problems, these three specific lines received particular attention on cardinal problems, thus suggesting that cardinal problems emphasized the importance of *Part 2* and *Part 3* to find the solution. This result supports the idea that cardinal quantities evoke a set-based representation.

Fourth, the analysis of backward eye movements from one line to a previous one informed us with regards to the information that participants came back to when reading the problems. In accordance with the visit duration analysis, participants made more regressions to lines 4 (*Part 3*) as well as to lines 2 and 5 (*Part 2*) on cardinal problems, as compared to ordinal problems. This indicates that participants' strategy includes looking back to previous lines for information about *Part 2* and *Part 3*, thus supporting the idea that participants were actively trying to find the value of *Whole 2* by adding up the (missing) values of *Part 2* and *Part 3*. This analysis confirms that participants tend to look back at information regarding *Part 2* and *Part 3* more often on cardinal problems.

Finally, a fifth measure provided new insights regarding our hypothesis that solving a problem whose initial representation is semantically incongruent with its solving algorithm requires to engage in a costly semantic recoding process to construct a new representation compatible with the available algorithm. Since pupil dilation varies closely in response to changes in task demands, pupillometry can be used as an indirect measure of participants' effort. By studying pupil dilation variations between success and failures on cardinal and on ordinal problems, we were able to measure how the cognitive load varied between situations. We predicted that participants' engagement in a semantic recoding step would result in an increase in pupil diameter on successfully solved cardinal problems. The results supported this hypothesis, since there was an increase in pupil diameter on successfully solved cardinal problems as compared to erroneously rejected cardinal problems. In other words, pupil dilation indicated an increased effort when participants managed to overcome their initial, incongruent representation of the problems and to find the solution to the cardinal problems. On the other hand, the pupil diameter difference between successes and failures on ordinal problems was not statistically significant. This can either indicate that there was no such difference since no semantic recoding was needed on ordinal problems, or it can simply be the sign of a lack of statistical power, since failures on ordinal problems were relatively scarce. Although we do not have the means to arbitrate between these two candidate explanations, the fact that there remained a significant difference between pupil dilation on correctly solved cardinal problems and on correctly solved ordinal problems seems to tip the scale in favor of the first interpretation. Indeed, it appears that correctly solving a cardinal problem required more effort, on average, than correctly solving an ordinal problem, which could be a sign of the existence of the hypothesized semantic recoding process.

Overall, our results support the SECO model according to which general, non-mathematical knowledge about the entities featured in a problem directly influences the representations that are constructed by the solvers, as well as their ability to find the solution. By showing an increased focus on subsets on cardinal problems, the use of eye-tracking helped support the idea that set-based representations are constructed whenever weights, prices, or collections are mentioned. Additionally, pupil dilation analysis confirmed that a costly semantic recoding process may be performed to overcome an incongruent representation and find the solution to a semantically incongruent problem. By furthering our understanding of the process of semantic recoding, we hope to identify ways to help students generate transfer between superficially dissimilar situations sharing a deeper bond.

References

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.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

Colomé, À., & Noël, M. P. (2012). One first? Acquisition of the cardinal and ordinal uses of numbers in preschoolers. *Journal of Experimental Child Psychology*, 113(2), 233-247.

Curtis, E. T., Huebner, M. G., & LeFevre, J. A. (2016). The relationship between problem size and fixation patterns

during addition, subtraction, multiplication, and division. *Journal of numerical cognition, 2(2)*, 91-115.

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H. C. (2015). Word problems: a review of linguistic and numerical factors contributing to their difficulty. *Frontiers in psychology*, *6*, 348.

De Corte, E., & Verschaffel, L. (1986). Eye-Movement Data as Access to Solution Processes of Elementary Addition and Subtraction Problems. *Annual Meeting of the American Educational Research Association*, San Francisco, April 16-20, 1986.

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.

Dewolf, T., Van Dooren, W., Hermens, F., & Verschaffel, L. (2015). Do students attend to representational illustrations of non-standard mathematical word problems, and, if so, how helpful are they?. *Instructional Science*, *43*(1), 147-171.

Fuson, K. C. (1988). *Children's counting and concepts of number*. New York: Springer-Verlag.

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.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*(2), 90-95.

Gros, H., Sander, E., & Thibaut, J. P. (2019). When masters of abstraction run into a concrete wall: Experts failing arithmetic word problems. *Psychonomic bulletin & review*, online first.

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.

Gros, H., Thibaut, J. P., & Sander, E. (2020a). Semantic Congruence in Arithmetic: A New Model for Word Problem Solving, Educational Psychologist, *in press*.

Gros, H., Thibaut, J. P., & Sander, E. (2020b). What we count dictates how we count: A tale of two encodings. *PsyArxiv*. doi: 10.31234/osf.io/4jev5.

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, *84*(1), 76-84.

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of educational psychology*, *87*(1), 18-32.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Cambridge, MA: Harvard University Press.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*(1), 109-129.

Lai, M. L., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., Lee, S. W. Y., ... & Tsai, C. C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, 10, 90-115.

Lyons, I. M., Vogel, S. E., & Ansari, D. (2016). On the ordinality of numbers: A review of neural and behavioral studies. In *Progress in brain research*, 227, 187-221.

Merkley, R., & Ansari, D. (2010). Using eye tracking to study numerical cognition: the case of the ratio effect. *Experimental Brain Research*, *206*(4), 455-460.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans and understanding: An inquiry into human knowledge structures*. New York, NY: Hillsdale.

Staub, F. C., & Reusser, K. (1995). The role of presentational structures in understanding and solving mathematical word problems. In C. A. Weaver III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 285-305). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Strohmaier, A. R., Tatsidou, K., Reiss, K. (2018). Eye movements during the reading of word problems. Advances in the use of eye tracking data. Fachgruppe Didaktik der Mathematik der Universität Paderborn (Hrsg.), *Beiträge zum Mathematikunterricht 2018* (S. 1759-1762). Münster: WTM-Verlag.

Thevenot, C. (2017). Arithmetic Word Problem Solving: The Role of Prior Knowledge. In Geary, D. C., Berch, D. B., Ochsendorf, J., Mann-Koepke, K. (Eds.), *Acquisition of Complex Arithmetic Skills and Higher-Order Mathematics Concepts* (pp. 47-66). Academic Press.

Thevenot, C., & Barrouillet, P. (2015). Arithmetic word problem solving and mental representations. *The Oxford handbook of numerical cognition*, 158-179.

Van der Schoot, M., Bakker Arkema, A. H., Horsley, T. M., & van Lieshout, E. C. (2009). The consistency effect depends on markedness in less successful but not successful problem solvers: An eye movement study in primary school children. *Contemporary Educational Psychology*, *34*(1), 58-66.

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*(1), 85-94.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Lisse: Swets & Zeitlinger.

Wasner, M., Moeller, K., Fischer, M. H., & Nuerk, H. C. (2015). Related but not the same: Ordinality, cardinality and 1-to-1 correspondence in finger-based numerical representations. *Journal of Cognitive Psychology*, *27*(4), 426-441.