

# Revealing Mental Representations of Arithmetic Word Problems Through False Memories: New Insights Into Semantic Congruence

Hippolyte Gros<sup>1</sup>, Jean-Pierre Thibaut<sup>2</sup>, Lucas Raynal<sup>3</sup>, and Emmanuel Sander<sup>3</sup>

<sup>1</sup> Centre de Recherche en Psychologie et Neurosciences, Aix-Marseille Université

<sup>2</sup> Laboratoire de l'Etude de l'Apprentissage et du Développement, Department of Psychology, Université de Bourgogne

<sup>3</sup> Instruction, Development, Education, and knowledge Acquisition Lab, Faculty of Psychology and Educational Sciences, University of Geneva

What can false memories tell us about the structure of mental representations of arithmetic word problems? The semantic congruence model describes the central role of world semantics in the encoding, recoding, and solving of these problems. We propose to use memory tasks to evaluate key predictions of the semantic congruence model regarding the representations constructed when solving arithmetic word problems. We designed isomorphic word problems differing only by the world semantics imbued in their problem statement. Half the problems featured quantities (durations, heights, elevator floors) promoting an ordinal encoding, and the other half used quantities (weights, prices, collections) promoting a cardinal encoding. Across three experiments, in French and in English, we used surprise memory tasks to investigate adults' mental representations when solving the problems. After the first solving task, the participants were given an unexpected task: either to recall the problems (Experiments 1 and 2) or to identify, from memory, the experimenter-induced changes in target problem sentences (Experiment 3). Crucially, all problems included a specific mathematical relationship that was not explicit in the problem statement and that could only be inferred from an ordinal encoding. We used the presence or absence of this relationship in the participants' responses to infer the structure of their representations. Converging results from all three experiments bring new evidence of the role of semantic congruence in arithmetic reasoning, new insights into the relevance of the cardinal–ordinal distinction in numerical cognition, and a new perspective on the use of memory tasks to investigate variations in the representations of mathematical word problems.

**Keywords:** arithmetic reasoning, recall, semantic encoding, mental models, ordinality


A central question in the arithmetic word problem-solving literature regards the nature of the problem representations constructed in working memory. Be it through the implementation of problem schemata akin to behavioral scripts (Kintsch & Greeno, 1985; Schank & Abelson, 1977), the construction of mental models depicting the problem situation (Johnson-Laird, 1983; Staub & Reusser, 1995; Thevenot & Barrouillet, 2015), or the abstraction of an interpreted structure describing the solvers' understanding of a given problem statement (Bassok, 2001), different theories have attempted to model the representational aspects of arithmetic word

problem solving. From an educational perspective, understanding the nature and the determinants of these mental representations is a key step in designing optimal school curricula for problem-solving education (Daroczy et al., 2015, 2020; Verschaffel et al., 2020).

In this article, we empirically assess central predictions of a recent model describing the representations of arithmetic word problems: the semantic congruence (SECO) model (Gros et al., 2020). In doing so, we pursue three complementary objectives. (1) First, we intend to assess the relevance of SECO in predicting the encoding, recoding, and solving of arithmetic word problems. (2) Second, we

This article was published Online First August 29, 2024.

Jennifer Wiley served as action editor.

Hippolyte Gros  <https://orcid.org/0000-0002-4151-0715>

This work was supported by grants from the CY Initiative (Grant CYIn-AAP2021-AmbEm-0000000026) awarded to Hippolyte Gros, the Regional Council of Burgundy, Paris Feder (Grants 20159201AAO050S02982 and 20169201AAO050S01845), awarded to Jean-Pierre Thibaut, and the Experimental Fund for the Youth and French Ministry of Education (Grant HAP10-CRE-EXPE-S1) and the French Ministry of Education and Future Investment Plan (Grant CS-032-15-836-ARITHM-0) awarded to Emmanuel Sander. The authors have no conflicts of interest to disclose.

This research was approved by the ethics committee of the University of Geneva (decision no. PSE.20181104.18). All participants gave informed consent to participate in the study. This work was not preregistered. The data for all three experiments are available on the Open Science Framework repository at <https://osf.io/5nqev/>

[view\\_only=6dcf3a21a2c840c6a16dce2bbf419762](https://doi.org/10.1037/xlm0001373) (Gros et al., 2024b).

Hippolyte Gros played a lead role in data curation, formal analysis, investigation, methodology, visualization, and writing—original draft and an equal role in conceptualization, funding acquisition, resources, and writing—review and editing. Jean-Pierre Thibaut played a supporting role in data curation, formal analysis, investigation, and methodology and an equal role in conceptualization, funding acquisition, resources, supervision, and writing—review and editing. Lucas Raynal played a supporting role in conceptualization, formal analysis, and investigation and an equal role in writing—review and editing. Emmanuel Sander played a supporting role in data curation, formal analysis, investigation, and methodology and an equal role in conceptualization, funding acquisition, resources, supervision, and writing—review and editing.

Correspondence concerning this article should be addressed to Hippolyte Gros, Centre de Recherche en Psychologie et Neurosciences, Aix-Marseille Université, UMR 7077, Marseille, France. Email: [hippolyte.gros@cri-paris.org](mailto:hippolyte.gros@cri-paris.org)

aim to evaluate the claim made by a recent line of studies in numerical cognition regarding the role played by the cardinal-ordinal distinction on adults' representations (Gros et al., 2021). (3) Third, we mean to explore the extent to which false memories can be used as a window into solvers' problem representations, thanks to an experimental design based on problem recall and sentence-recognition tasks.

### Inside the Semantic Congruence Model

The SECO model proposes that when attempting to solve a word problem, the problem statement is abstracted into an interpreted structure that depends both on the world semantics and on the mathematical semantics evoked by its wording (see Figure 1). In other words, solvers construct a mental representation that does not solely depend on the mathematical knowledge relevant to the problem (the mathematical semantics) but also on the nonmathematical knowledge about the world (the world semantics) that is related to the entities mentioned in the problem statement. This interpreted structure may either be specified into a solving algorithm when available, or it may, through a costly recoding process, be replaced by another representation, closer to the deep structure of the problem. This new, more abstract representation would then make it possible to use a different solving algorithm (Gros et al., 2020).

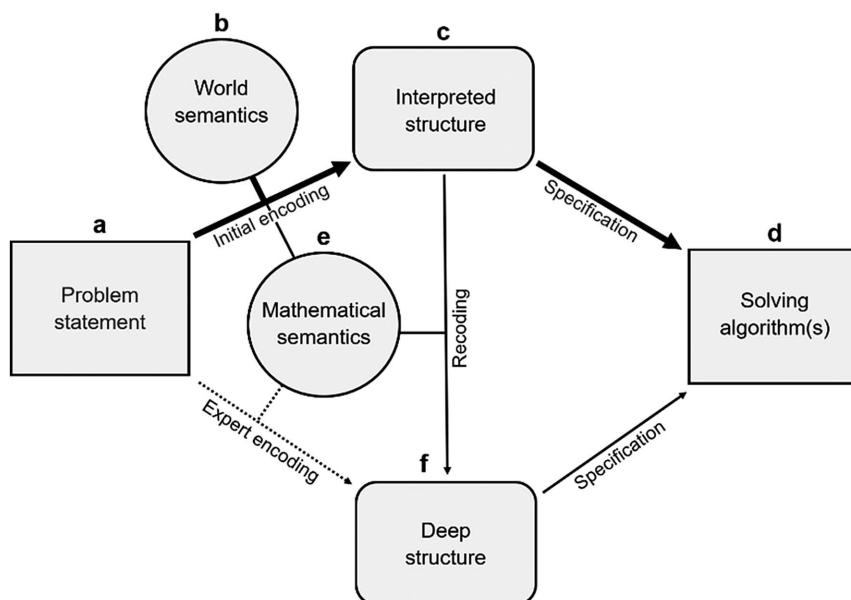
According to SECO, experts might also directly abstract a problem's deep structure, completely bypassing the world semantics attached to the problem statement. This alternative pathway was inspired by previous works suggesting that experts may see beyond the surface features of a problem statement and focus on its deep structure, contrarily to novices (see Chi et al.'s, 1981, seminal study on the categorization criteria guiding experts and novices).

However, it should be noted that even experts may struggle with using this direct encoding pathway at times, as recent evidence indicates they can still be influenced, to a degree, by world semantics (Gros et al., 2019).

SECO notably differs from previous frameworks in two main respects: (a) its description of the processes at play in the initial encoding of the problems and (b) its proposal for the existence of a costly recoding pathway making it possible to go beyond this initial representation. First, regarding the initial encoding of the problems, the schema framework (Kintsch & Greeno, 1985) predicts that solvers identify key terms in the propositional structure of the problem statement to retrieve the relevant solving schemata from memory and implement them with the problems' numerical values to find the solution. According to this perspective, problems sharing the same propositional structure (for instance, a problem mentioning "John has three apples less than Emily" and one stating "The English class lasted 3 hr less than the chemistry class") should both cue the use of the same difference schema describing a three-unit difference between the two quantities and, thus, lead to the same solving algorithm. This differs from SECO's proposition that solvers' knowledge and experience about apples or chemistry classes will influence the structure of the mental representation they will construct of the situation and, thus, the solving algorithm they will use.

Alternatively, the situation model framework (Johnson-Laird, 2010; Reusser, 1990) posits that solvers, upon encountering a problem statement, construct a unique episodic situation model tailored to that specific problem (Reusser, 1990). According to this framework, the constructed situation model encompasses all functional relationships detailed in the problem statement, suggesting an idiosyncratic interpretation for each problem. Unlike the schema

**Figure 1**  
*The SECO Model*



*Note.* SECO = Semantic Congruence. Reprinted from "Semantic Congruence in Arithmetic: A New Conceptual Model for Word Problem Solving" by H. Gros, J.-P., Thibaut, E. Sander, 2020, *Educational Psychologist*, 55, p. 5 (<https://doi.org/10.1080/00461520.2019.1691004>). Copyright 2020 by Taylor & Francis. Reprinted with permission.

approach, this framework allows for the possibility of encoding differences between problems that share the same mathematical structure but differ in the entities mentioned in their statements. However, it does not predict that those problem statements involving different entities may also lead to similar encodings when sharing the same aspects of world semantics. In other words, while it predicts that different situations may lead to different encodings, it says nothing about the potential regularities observed across those encodings, based on the world semantics imbued in the problem statements. SECO, on the other hand, predicts that structural consistencies may emerge across different interpreted structures, depending on the semantic dimensions influencing the initial encoding of the problem statements. According to SECO, the role played by world semantics may lead to commonalities in how problems are understood and represented. This prediction means that by using isomorphic problems sharing the same mathematical structure but referring to distinct aspects of world semantics, it should be possible to lead to distinct representations being abstracted and, thus, to specific solving strategies being used.

Second, regarding the potential existence of a recoding pathway, the competing theories say little about this eventuality. While the semantic alignment framework (Bassok, 2001) proposes that solvers will struggle with solving problems whose interpreted structure is “misaligned” with their solution, it does not address the possibility that solvers may overcome this obstacle. SECO, however, predicts that solvers may attempt to recode their initial encoding when it fails to lead to adequate or optimal solving strategies. Such semantic incongruence between the interpreted structure and the targeted solving algorithm may be overcome by engaging in a semantic recoding of the constructed representation (the recoding pathway from *c* to *f* in Figure 1). According to this perspective, solvers are not necessarily confined to their initial representation of the problems. Instead, they may engage in a cognitively demanding recoding to construct a new representation that will make it possible to use alternative solving strategies. In other words, even on a given problem statement, different consecutive representations may be abstracted, potentially leading to different strategies being used.

This article’s first objective (1) is to assess the validity of these two predictions regarding (a) the commonalities found across different interpreted structures, depending on the semantic dimensions influencing the initial encoding of the problem statements, and (b) the potential for solvers to engage in a costly recoding process, enabling them to go beyond their initial representation. To achieve this, we will investigate the influence of world semantics on the encoding of cardinality and ordinality in arithmetic word problems.

### The Case of the Cardinal–Ordinal Distinction: A Look Into Semantic Determinants of Problem Solving

As stated above, the second objective (2) is to collect new data to further characterize the importance of the cardinal–ordinal distinction on adults’ mathematical reasoning. This goal stems from recent works suggesting that numerical situations may be encoded either into a representation emphasizing the cardinal properties of numbers (i.e., a cardinal encoding) or into a representation emphasizing their ordinal properties instead (i.e., an ordinal encoding; Gros et al., 2019, 2021).

The notions of ordinality and cardinality express two sides of numbers: ordinality refers to their existence as an item in an ordered

list, while cardinality refers to their meaning as the total number of entities being counted. This distinction is fundamental in mathematics (Dantzig, 1945; Frege, 1980; Russell, 1919), especially in set theory (Dauben, 1990; Suppes, 1972), and several works in developmental psychology have shown that it has implications beyond formal mathematics. For instance, Gelman and Gallistel (1986) argued that in learning how to count, children need to understand both the ordinal and the cardinal properties of numbers. On the one hand, the “stable-order principle” refers to children learning that the list of words used to count needs to be said in a definite and stable order, each word having the same predecessor and the same successor over trials. On the other hand, the “cardinal principle” refers to the understanding that the final word of an enumeration indicates the total number of entities in the set being counted. More recently, a growing number of studies have investigated the development of the cardinal meaning of numbers (e.g., Bermejo, 1996; Geary & vanMarle, 2018; Le Corre & Carey, 2007; Sarnecka & Lee, 2009; Shusterman et al., 2016; Wynn, 1992), the development of their ordinal meaning (Cheung & Lourenco, 2019; Fischer & Beckey, 1990; Hund et al., 2021; K. Miller et al., 2000, 2015), as well as the differences between the developmental trajectories of these two notions (Baccaglini-Frank et al., 2020; Colomé & Noël, 2012; Meyer et al., 2016; Wasner et al., 2015).

In the field of arithmetic problem solving, Verschaffel et al. (1999) showed that additive problems dealing with the ordinality of numbers presented specific challenges to upper elementary school pupils. Since then, a growing body of research has shown that the perception of cardinality and ordinality retains a significant role when conceiving of numerical situations: Even after counting procedures are acquired, there remains an ontological difference between the way adults conceive of numbers either as order labels or as count values (Gamo et al., 2010; Gros et al., 2019, 2021).

This difference was empirically investigated by using problems sharing the same mathematical structure but allowing two distinct solving strategies (Gros et al., 2021). It was hypothesized that depending on the solvers’ encoding of the problems (the interpreted structure), participants would preferentially use one of the two available solving strategies. The key aspect of this study was that depending on the world semantics imbued in the problems, it was predicted that the solvers would preferentially use one of the two different encodings. Consider, for instance, the following problem:

Paul has 8 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 3 green marbles less than Paul has red marbles. How many marbles does Jolene have?

Because this problem involves counting collections of marbles, which have no intrinsic order to them, it was hypothesized that such a problem would emphasize the cardinal nature of the numbers it features (Gros et al., 2021). Since there is no reason to mentally line up the marbles in a specific order, participants tend to think of the marbles of different colors as distinct, autonomous entities organized as subsets to be combined. Therefore, when asked to solve the problems, participants attempt to calculate the number of marbles Jolene has by looking for the number of blue marbles she has and adding it to the number of green marbles she has. That is, most participants use a three-step strategy to solve this problem:  $14 - 8 = 6$ ;  $8 - 3 = 5$ ;  $6 + 5 = 11$  (Gros et al., 2021, 2024a).

They identify that Jolene has six blue marbles and five green marbles, thus adding up to 11 marbles in total.

On the other hand, consider the following duration problem:

The construction of the palace took 8 years. Plans for the construction were made beforehand. The construction of the palace was completed in year 14. The construction of the castle started at the same time as the construction of the palace. The construction of the castle took 3 years less than the construction of the palace. When was the construction of the castle completed?

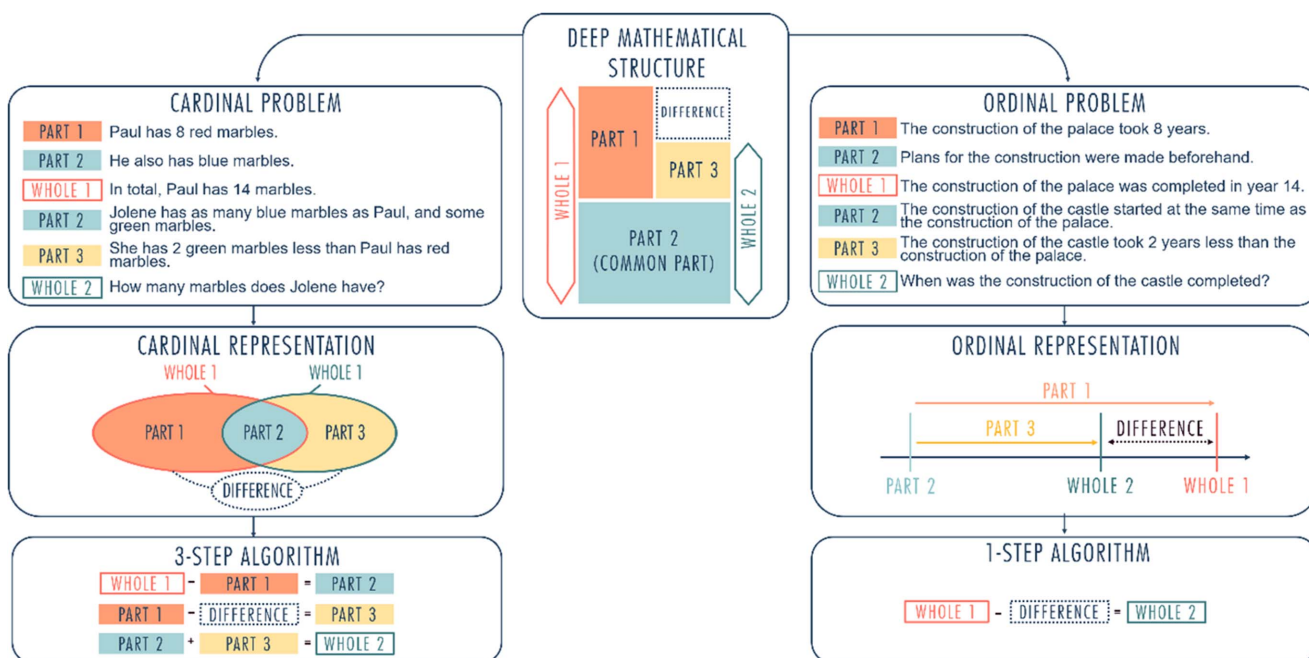
This problem has the same mathematical structure as the marble problem, but because it involves duration values instead of marble counts, it was hypothesized that participants would tend to conceive of the described situation as being ordered along a timeline. Due to solvers' nonmathematical knowledge about durations, the different entities in the problems are not perceived as parts and wholes but as states (positions on a timeline) and changes (events with a duration) (Gros et al., 2021). This ordinal representation thus features two changes (the palace and the castle's respective construction times) starting from the same state, that is, the same starting point on the timeline. By encoding these two transitions along a time axis, it becomes clear that the difference between their respective lengths is equal to the difference between their respective endpoints. In other words, since both constructions started at the same time, the difference between the time to build the castle and the time to build the palace is equal to the difference between the time at which the castle was completed and the time at which the palace was completed. This ordinal representation makes it possible for the participants to understand that there is a much shorter solving strategy to be found:  $14 - 3 = 11$ .

The same one-step strategy could have also been used to solve the marble problem, but participants rarely used it (Gros et al., 2021). Indeed, using the one-step strategy on the marble problem would require understanding that the difference between Paul's green marbles and Jolene's red marbles is equal to the difference between Paul's total number of marbles and Jolene's total number of marbles. In other words, since Paul and Jolene both have the same number of blue marbles, there is no need to calculate this number nor to calculate how many green marbles Jolene has. Instead, one only needs to infer that since Jolene has three green marbles less than Paul has red marbles, and since they both have the same number of blue marbles, then Jolene simply has three marbles less than Paul in total (Gros et al., 2021).

On the duration problem, on the other hand, it was argued that an ordinal, timeline-based representation allows for a direct comparison of the time it took to build the palace and the time it took to build the castle (Gros et al., 2021). The salience of this comparison would thus make it easier to infer that since both constructions started at the same time, and since the construction of the castle took 3 years less than the construction of the palace, then the construction of the castle was completed 3 years before the construction of the palace. In other words, in both problems, it is unnecessary to calculate the value of the *common part* to find the solution. Figure 2 describes the deep mathematical structure shared by these two problems, as well as the representations that are thought to be abstracted, resulting in different solving strategies being used. In order to use the one-step strategy "*Whole 1 - Difference = Whole 2*," one needs to understand that the difference described between *Part 1* and *Part 3* is equal to the difference between *Whole 1* and *Whole 2*. According to the

**Figure 2**

*Graphical Summary of the Encoding Differences Between Cardinal and Ordinal Problems*



*Note.* See the online article for the color version of this figure.



solving strategies used by participants, this inference is significantly easier to make on ordinal than on cardinal problems (Gros et al., 2021), to the point that even expert mathematicians would struggle to make this inference on cardinal problems (Gros et al., 2019).

In Gros et al. (2021), it was proposed that the mention of specific quantities in a problem statement can trigger a particular encoding of the situation. Quantities usually associated with cardinality, such as weight problems, price problems, and collection problems, are supposed to induce a cardinal encoding. On the other hand, quantities whose world semantics highlight ordinality, such as duration problems, height problems, and elevator problems, are hypothesized to evoke a cardinal encoding. Empirically, this difference between cardinal and ordinal quantities proved impactful not only for strategy choice but also in problem comparison, problem sorting, solvability assessment, strategy transfer, and solution evaluation tasks (Gros et al., 2019, 2021). However, directly probing the very representations underlying these differences is not a straightforward task, since the mental representation itself cannot be measured. In fact, all previous measures of these representational differences relied on participants' ability to find a correct solving strategy. Here, we propose to take a closer look at an intermediate step between the initial reading of the problem and the writing down of its solution. We investigate the inferences drawn from the problem encoding itself and evaluate how they are linked to the world semantics in the problem statement. We intend to show that the encoding of ordinal problems leads to a specific mathematical inference that will result in false memories and predict the solving strategy used by the participants.

### Probing Representations With Recall and Recognition Tasks

Since no means of direct investigation of the representations themselves are available, text recall tasks and sentence-recognition tasks can bring valuable information regarding the nature of the encoded representations: If there is indeed a crucial difference between one's understanding of ordinal and cardinal situations, then this difference should result in different encodings being constructed and memorized. We strove to assess the validity of this claim by evaluating the presence of specific inferences that can only be drawn from the problem statement if participants encode an ordinal representation of the situation. We used these inferences to assess whether participants constructed and memorized an ordinal representation of the situation. Indeed, previous works on text comprehension suggest that sentences from which inferences can be drawn may mislead participants both in recognition (Bransford & Johnson, 1973; Kintsch & Bates, 1977; Kintsch et al., 1990; Noordman & Vonk, 2015) and in recall tasks (Black & Bern, 1981; Corbett & Doshier, 1978; Kintsch & Van Dijk, 1978; Sulin & Dooling, 1974). In Bower et al.'s (1979) seminal work on the importance of scripts in text comprehension, it was shown that participants tended to infer actions that were not explicitly described in the text but that were coherent with the scenario depicted by the text. Those inferences led participants to erroneously recall events that were never described in the text, with a surprisingly high degree of confidence in their recalls. Here, we predicted that the implicit inferences drawn from one of the two possible encodings of the problems would lead participants to erroneously remember pieces of information which were not initially present in the problems but

which could be easily inferred from an ordinal representation of the situation.

Our work builds on previous paradigms using recognition tasks to evaluate which inferences were included in participants' representation of a given situation. For instance, in one of their experiments, Mani and Johnson-Laird (1982) designed a task to investigate participants' mental representation of spatial descriptions. They presented participants with four-sentence descriptions of spatial configurations of the form "A is to the left of B. C is to the right of B. D is in front of A. E is in front of B." Participants had to evaluate whether specific diagrams were consistent with the spatial description. After the task, participants were presented with an unexpected recognition task in which they had to identify among a series of four-sentence statements which were the ones they had been presented before and which were new, previously unseen statements. The authors used three types of test statements: Some were identical to the ones previously seen, some described different spatial configurations from the ones previously seen, and some described a spatial configuration that was inferable from the previously seen statements, but which differed from the propositional structure of the initial spatial description. Interestingly, the authors showed that the statements presenting an inferable spatial configuration tended to be erroneously recognized more often than the ones presenting a different spatial configuration. In other words, participants tended to base their recognition of the texts on the mental model they had constructed of the problems, instead of only using the verbatim text as a basis for recognition. This property of text recognition is especially relevant to our study, as it indicates that misremembrance can inform on the representations constructed by the participants.

In a similar spirit, Verschaffel (1994) used a solving task followed by a retelling task to investigate the nature of the representations constructed by fifth graders when solving arithmetic word problems. Specifically, he looked at problems whose wording was consistent with the arithmetic operation needed (e.g., using "has X more than" for addition problems, or "has Y less than" for subtraction problems) versus problems where the wording was inconsistent with the operation (e.g., addition problems mentioning "has Z less than"). He found that when the wording was not consistent with the operation needed to solve the problem, children often mistakenly changed the wording when they explained the problem back, making it consistent with the operation. This was interpreted as a clear indicator that children's mental representation of the problems differed from their exact wording. Hegarty et al. (1995) made another attempt to investigate the representations constructed by individuals engaged in arithmetic word problems solving. They conducted a problem-solving task followed by a text recall task and a text recognition task to evaluate which participants constructed a problem model. They showed that the more successful solvers tend to make a lower number of semantic mistakes but a higher number of literal mistakes than the less successful solvers. According to the authors, this finding was consistent with the hypothesis that the more proficient participants had constructed a problem model, since they recalled the semantic structure of the problems successfully but were less accurate in recalling the exact wording of the problems. On the other hand, the less proficient participants may not have constructed a problem model, since they recalled the problems' wording accurately but made more mistakes with regard to their semantic structure.

In the field of arithmetic word problem solving, Thevenot (2010) used a text recognition task to investigate some of the factors influencing participants' mental representations of the problems. The author asked participants to solve a series of problems and then presented them with an unexpected recognition task. Ingeniously, she used three different types of problems in the recognition task: (a) problems identical to the original problems, (b) new problems mathematically inconsistent with the original ones despite differing by only one or two words, and (c) paraphrastic problems mathematically consistent with the original problems despite differing by a total of three words. Results revealed that the paraphrastic problems were more often erroneously recognized than were the new inconsistent problems, despite the latter differing by a lower number of words from the original problems. This indicated that participants had constructed and memorized a representation of the problems that depended on the structure of the situations they described rather than on their precise wording.

Building upon these previous studies on text recognition and text recall and in accordance with the third objective (3) we set for this study, we intend to go one step further and use both recall tasks and sentence-recognition tasks to contrast and compare different representations of problems sharing the same mathematical structure. By using erroneous recollections as a source of insights into solvers' encoding of arithmetic word problems, we aim to demonstrate a link between the world semantics imbued in the problems and the presence of specific inferences in the constructed representations. The errors made by the participants in retelling or recognizing the problem statement should make it possible to take a closer look at the encoding and recoding processes described in the SECO model.

## The Present Study

As previously stated, this article pursues a threefold objective: (1) to evaluate key predictions of the SECO model, (2) to better understand the encoding of cardinality and ordinality in arithmetic word problems, and (3) to show how false memories may help differentiate between different encodings of a given problem. Across three experiments, we presented participants with a solving task followed by an unexpected task testing their recollection of the problems' statements. In the first two experiments, the solving task was followed by an unexpected recall task in which participants had to retell from memory the statements of the problems they just solved. Due to additional information being inferred from ordinal representations but not from cardinal representations, we expected participants to erroneously recall specific inferences in the retell task on ordinal problems but not on cardinal problems. In a third experiment, we presented participants with an unexpected sentence-recognition task, in which they had to decide whether target sentences were taken from the problems they previously solved, or whether these sentences had not been included in the problem statements. The modified sentences clearly stated certain mathematical relationships that were not explicit in the initial statements. We expected participants to be more likely to erroneously recognize the modified sentences of ordinal problems than that of cardinal problems. The first two experiments were conducted in French, the third one in English, which decreases the probability that the observed phenomenon would stem from idiosyncratic properties of the French or English language.

## Experiment 1

Experiment 1 was a first attempt to use a recall task to gather evidence regarding the inferences that can be drawn from an ordinal representation but not from a cardinal representation. Using SECO, we formulated three hypotheses regarding the solving task and the problem recall task. (Hypothesis 1) First, during the solving task, we predicted that we could replicate the results from Gros et al. (2021), meaning that due to the different representations of the problems, the participants' ability to use the shortest strategy would depend upon the cardinal versus ordinal nature of the problems: Participants should be more prone to using the one-step strategy on duration problems than on collection problems, in accordance with SECO's predictions.

(Hypothesis 2) Second, based on SECO's depiction of the role of world semantics on the initial encoding of arithmetic word problems, we anticipated that problems mentioning ordinal quantities would lead to the encoding of an ordinal interpreted structure, while problems mentioning cardinal quantities should initially lead to the encoding of a cardinal interpreted structure. Thus, regarding the unexpected recall task, we predicted that if participants did construct an ordinal encoding of the situation described in the problems featuring ordinal quantities, then they would be more likely to make a specific inference regarding the nature of the difference mentioned in the problems. More precisely, while the fifth sentence of the problems always introduced a difference between *Part 3* and *Part 1* (e.g., "Jolene has two green marbles less than Paul has red marbles" in a cardinal problem or "The construction of the castle took 2 years less than the construction of the palace" in an ordinal problem; see Figure 2), we predicted that participants' recollection of the fifth sentence would differ between cardinal and ordinal problems. On ordinal problems, we predicted that participants' ordinal representation would sometimes lead them to misremember the difference between *Part 3* and *Part 1* as a difference between *Whole 2* and *Whole 1* instead, since they are hypothesized to have included this inference into their representation. In other words, instead of stating that "The construction of the castle took 2 years less than the construction of the palace," they might state that "The construction of the castle was completed 2 years before the construction of the palace." On the other hand, cardinal problems should not lead to this error, since the inference that the *Part 3–Part 1* difference is equal to the *Whole 2–Whole 1* difference is much harder to make when reasoning with a cardinal encoding. Participants would thus be more likely to make this specific *whole-to-whole error* while retelling ordinal problems (e.g., "The construction of the castle was completed 2 years before the construction of the palace"), than while retelling cardinal problems (e.g., "Jolene has two marbles less than Paul").

(Hypothesis 3) Finally, based on SECO's prediction that the problem representation encoded by a solver—either their initial interpreted structure or its later recoding—determines the solving strategy they use, we anticipated that participants' solving strategies would be linked to their encoded representations, regardless of the cardinal versus ordinal semantics imbued in the problem statements. Thus, we made the hypothesis that the aforementioned recall mistake would be more likely to occur on the problems that were solved using the one-step strategy, than on those solved using the three-step strategy. Indeed, the use of the one-step strategy is thought to be linked with an ordinal encoding of the situation (either as an initial encoding or after a recoding process), so solving a

problem in one-step, rather than in three steps, should be associated with a higher chance of whole-to-whole inferences being remembered in place of the part-to-part inferences.

## Method

### Participants

We estimated a sample size using the BUCSS R package (Anderson & Kelley, 2018), based on results from a previous study using similar materials (Experiment 4; Gros et al., 2021). After correction for uncertainty and publication bias following Anderson et al.'s recommendations (2017) and using a high level of targeted statistical power (0.95), we estimated a minimum sample size of 45 for the solving task. However, since we had no previous results on participants' propensity to spontaneously produce the specific recall error targeted by our experiment, we elected to double this sample size for the recollection task, to account for potentially lower effect sizes. Thus, we set to recruit at least 90 participants. After distributing the survey online, through social networks and mailing lists, we waited 2 weeks before closing data collection. A total of 140 people had participated at this time. Among them, 13 left at least one of the questions unanswered and were subsequently excluded from our data set. The analyses were conducted on the remaining 127 participants (81 women,  $M_{\text{age}} = 33.39$  years,  $SD = 8.15$ ). All participants voluntarily took part in the experiment, without any monetary incentive. They all indicated speaking French fluently.

### Materials

The materials were based on previous works focusing on the difference between cardinal and ordinal encodings (Gros et al., 2021). To maximize the encoding difference between the problems, we selected our materials from the two most stereotypical quantities used in Gros et al. (2021): collection problems (see Table 1, column A) and duration problems (see Table 1, column B). We used a

within-subject design to allow for within-subject comparisons between responses on cardinal and on ordinal problems. Each participant was presented with one randomly chosen cardinal problem (a collection problem) and one randomly chosen ordinal problem (a duration problem) (see Table A1 for the original problem statements, in French). All the problems were isomorphic, and the numerical values used were randomized across problems, according to the following rule:  $15 \geq \text{Whole} > \text{Part} > 4 > \text{Difference} \geq 2$ .

### Procedure

This experiment was conducted online using the Qualtrics platform for online experiments. On the first page, the instructions read:

On the next page, you will find an arithmetic problem. Please take the time to read it carefully. Your task is to try to solve the problem using as few operations as possible. We ask that you take enough time to read and understand the problem, as this is not a speed test. Remember that the goal is to solve the problems using as few operations as possible. Type down every operation(s) that you used to come up with the solution, even the simplest one(s) that you can mentally calculate. For instance, the computation " $15 - 6 - 2 = 7$ " should not be written as a unique operation, but broken down as " $15 - 6 = 9$ " and " $9 - 2 = 7$ ," which then count for two operations. *Translated from French*

On the next page, a problem was presented, meant to evoke either a cardinal encoding (collection problem) or an ordinal encoding (duration problem). Along with the problem statement, the instructions to solve the problem using as few operations as possible were reminded to the participants. Below the problem statement, a text box allowed participants to write down the operation(s) they used, and another text box was used to write down the problem's solution. On the next page, the initial instructions were repeated, and the second problem followed. Depending on which problem participants had been presented initially, the following problem was chosen to evoke a different encoding. In other words, if participants had to solve a

**Table 1**

*The Four Cardinal and Ordinal Problem Statements Used in Experiment 1 and Experiment 2*

| Problem identifier | A. Cardinal problems   | Problem identifier | B. Ordinal problems   |
|--------------------|--|--------------------|---|
| Collection 1       | Paul has 5 red marbles.<br>He also has blue marbles.<br>In total, Paul has 14 marbles.<br>Jolene has as many blue marbles as Paul, and some green marbles.<br>She has 2 green marbles less than Paul has red marbles.<br>How many marbles does Jolene have?                            | Duration 1         | Sofia travelled 5 hours.<br>Her trip started during the day.<br>Sofia arrived at 14 h.<br>Fred left at the same time as Sofia.<br>Fred's trip lasted 2 hours less than Sofia's.<br>What time was it when Fred arrived?  |
| Collection 2       | Sarah owns 5 goldfish.<br>Her other pets are all iguanas.<br>In total, she owns 14 pets.<br>Bobby is pet-sitting Sarah's iguanas during the holidays, he puts them with his pet turtles.<br>Bobby owns 2 turtles less than Sarah owns goldfish.<br>How many pets are there at Bobby's? | Duration 2         | The construction of the palace took 5 years.<br>Plans for the construction were made beforehand.<br>The construction of the palace was completed in year 14.<br>The construction of the castle started at the same time as the construction of the palace.<br>The construction of the castle took 2 years less than the construction of the palace.<br>When was the construction of the castle completed? |

*Note.* Translated from French.

cardinal problem first, then the second problem was ordinal. Problem order was randomized between participants. When participants had solved both problems, they were presented with an unexpected recall task. They were told that they had to recall as precisely as possible the text from the first of the two problems. They were instructed to write everything they remembered about the problem statement as faithfully as possible. After they had completed this task, the next page asked them to write down the text of the second problem they had to solve.

The data for all three experiments are available on the Open Science Framework repository at [https://osf.io/5nqev/?view\\_only=6dcf3a21a2c840c6a16dce2bbf419762](https://osf.io/5nqev/?view_only=6dcf3a21a2c840c6a16dce2bbf419762).

## Results

First, participants' answers to the solving task were analyzed. In 95.21% of the cases, the strategies used by the participants to solve the problems could easily be inferred from their report of the operations they used to solve the problems. The 4.79% of cases where the strategy could not be directly inferred from their response (correct response provided with no operation leading to it) were classified as "unidentified" (see Figure 3).

The identifiable responses were either classified as "one-step strategy" (successful use of the shortest strategy), "three-step strategy" (successful use of the longest strategy), or "error" (wrong operations leading to a false answer). The distribution of the participants' solving strategies depending on the ordinal versus cardinal nature of the problems is described in Figure 3. We used a generalized linear mixed model with a binomial distribution to evaluate how likely participants were to use the one-step strategy on cardinal and on ordinal problems. We used the successful use of the

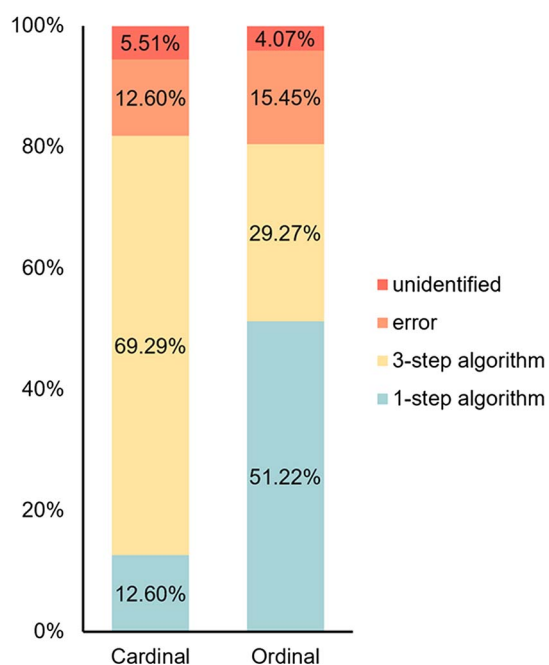
one-step strategy in the 254 recorded trials as the dependent variable, the semantic nature (cardinal vs. ordinal) of the problems as a fixed effect, and included participants as a random intercept (Variance = 2.52,  $SD = 1.59$ ). We performed all the analyses in this article using R (R Core Team, 2019) and lme4 (Bates et al., 2015). The model successfully converged, with a total explanatory power of 49.4% (conditional  $R^2$ ). As in Gros et al. (2021), the participants were considerably more likely to discover the one-step strategy on ordinal problems (51.22%) than on cardinal problems (12.60%), thus supporting Hypothesis 1, Estimate = 2.82,  $SE = 0.60$ ,  $z = 4.67$ ,  $p < .001$ .

Second, to test our second hypothesis (Hypothesis 2) regarding the encoded representation, we analyzed the problem statements they attempted to recall. Namely, we focused on the fifth sentence of the problems describing the difference between *Part 1* and *Part 3* (see Figure 2) and coded whether the participants had erroneously recalled the difference between *Part 1* and *Part 3* as a difference between *Whole 1* and *Whole 2* (a whole-to-whole recall mistake). In other words, for the marble problem, we counted how often participants recalled "Jolene has  $x$  marbles less than Paul has red marbles" (whole-to-whole difference) instead of the correct sentence "Jolene has  $x$  green marbles less than Paul" (part-to-part difference). For the duration problem, we counted how often participants recalled a sentence stating "The construction of the castle ended 3 years before that of the palace" (whole-to-whole difference) instead of the correct sentence "The construction of the castle took 3 years less than that of the palace" (part-to-part difference).

Table 2 indicates the distribution of participants' erroneous recalls of a whole-to-whole difference on cardinal and on ordinal problems. Participants recalled this inference instead of the proper problem phrasing in 15.75% of the ordinal problems, whereas they only made this mistake in 0.79% of the cardinal problems. Due to the extremely low number of participants who erroneously recalled a sentence indicating a whole-to-whole difference on a cardinal problem (1 out of 127), the comparison between cardinal and ordinal problems could not be done using variance analysis. Instead, we performed a McNemar test between these two conditions, which indicated, as hypothesized, that the semantic nature of the problems (ordinal vs. cardinal) had a significant impact on the rate of erroneous whole-to-whole sentences being recalled by participants,  $\chi^2(1, N = 127) = 15.43$ ,  $p < .001$ . Of note, it appears that erroneous recalls were more frequent for the Duration 2 problem (16 erroneous recalls) than on the Duration 1 problem (four erroneous recalls). This disparity might suggest that, in addition to the semantic differences between cardinal and ordinal problems, other factors could also influence the error rate in the recall task. However, given the overall low number of errors and the fact that each participant only had to solve one of the two ordinal problems, it was not feasible to investigate this potential difference further in this experiment.

Finally, to assess the link between strategies and encoded representations (Hypothesis 3), we investigated which strategy used to solve the ordinal problems in the solving task was the most likely to lead to an erroneous mention of the whole-to-whole inference in the recall task. Since there was only one occurrence of a participant referring to a whole-to-whole difference on a cardinal problem, we focused our analysis on the strategies used to solve ordinal problems. Crucially, most cases in which a participant erroneously recalled a sentence referring to a whole-to-whole difference

**Figure 3**  
Strategy Distribution in Experiment 1 Depending on the Quantities Used in the Problems



Note. See the online article for the color version of this figure.



**Table 2**

*Distribution of Participants' Erroneous Whole-to-Whole Recalls on Cardinal and Ordinal Problems*

| Problem type      | No mention of a whole-to-whole difference | Erroneous recall of a whole-to-whole difference |
|-------------------|---|---|
| Cardinal problems | 126 (99.21%)                              | 1 (0.79%)                                       |
| Ordinal problems  | 107 (84.25%)                              | 20 (15.75%)                                     |

regarded ordinal problems solved using the one-step strategy (75.00% of the recall mistakes), although a small portion of the whole-to-whole recall mistakes were linked either to answers that were insufficiently detailed to be interpreted (unidentified answers: 10.00%) or to errors in the solving task (15.00%) (see Table 3). Thus, while 29.27% of the ordinal problems were solved using the three-step strategy, none of the erroneous recollection of a whole-to-whole difference happened on such trials. Additionally, among the 63 ordinal problems solved in one step, 15 (23.81%) led to the erroneous recollection of a whole-to-whole difference, whereas there was no erroneous recollection in the 36 ordinal problems solved in three step.

## Discussion

This experiment provides insight into the problem representations constructed by the participants. We evaluated three hypotheses derived from the SECO model. First, in accordance with Hypothesis 1, the results corroborated Gros et al.'s (2021) original findings by showing that the choice of a solving strategy was directly dependent on the cardinal versus ordinal nature of the quantities used in the problem. That was a first indicator that different problem representations had been constructed, as predicted by SECO. Participants had been explicitly instructed to use the shortest strategy they could think of, using as few operations as possible, but only 12.60% of them managed to find the one-step strategy on cardinal problems, whereas more than half of the participants used this one-step strategy to solve the ordinal problems.

Second, and most importantly, the analysis of the recall mistakes made by the participants in the recall task supported Hypothesis 2 and provided new insights regarding the nature of the representations constructed by the participants (see Table 2). As hypothesized, participants were more likely to misremember the sentence describing the difference between *Part 1* and *Part 3* as a difference between *Whole 1* and *Whole 2* on ordinal problems (20 participants)

**Table 3**

*Strategies Leading to the Erroneous Recall of Whole-to-Whole Sentences on Ordinal Problems*

| Solving strategy    | No mention of a whole-to-whole difference | Erroneous recollection of a whole-to-whole difference |
|---------------------|---|---|
| One-step strategy   | 48  | 15  |
| Three-step strategy | 36  | 0   |
| Solving error       | 16  | 3   |
| Unidentified answer | 7   | 2   |

than on cardinal problems (only one participant). Despite being mathematically true, this piece of information regarding the difference between *Whole 1* and *Whole 2* was not directly present in the original problem statements; participants had to infer it from the situation described. The fact that participants were significantly more likely to assume that the difference was presented between *Whole 1* and *Whole 2* on the original ordinal problems indicates that the abstracted representation of ordinal problems made it more salient. In other words, it supports SECO's prediction (a) that the mention of durations within a problem elicits an ordinal, axis-based representation, making it easier to understand that if two events start at the same time and one is  $x$  years shorter than the other, then one ends  $x$  years before the other (see Figure 2). In accordance with SECO's prediction regarding the role of world semantics on the encoding of the problems, this inference appeared more frequently on duration problems than on collection problems.

Third, in line with Hypothesis 3, most of the participants who erroneously recalled the difference as a difference between wholes instead of a difference between parts in an ordinal problem did so after solving the problems using the one-step strategy (see Table 3). Indeed, none of the participants who had solved the ordinal problem using the three-step strategy did this mistake. Finding the solution in one step made it more likely to erroneously recall a sentence describing a whole-to-whole difference instead of the part-to-part difference. This corroborates the assumption that the misremembrance of the fifth sentence is a key indicator of the nature of the representations encoded by the participants. Overall, this experiment showed that the differences between cardinal and ordinal encodings tampered with participants' recollection of the problem statements. Participants falsely recalled sentences that were not present in the problems, due to the specific representation they had constructed. However, the task presented to the participants was relatively short and easy, and a rather low number of recall mistakes were made in the recall task. We considered whether making the task more difficult would increase the likeliness of witnessing an erroneous recall of the problems. In this perspective, we designed a second experiment in which we doubled the number of problems to solve and recall, in an effort to increase the task difficulty while assessing the replicability of the observed differences.

## Experiment 2

This second experiment attempted to build upon the first experiment and gather converging evidence in a different setting, using a higher number of problems to increase task difficulty. The three hypotheses were the same as in Experiment 1, since the goal was to identify whether participants would still recall a greater number of whole-to-whole inferences on ordinal than on cardinal problems.

## Method

### Participants

We estimated a sample size using the same rationale as in Experiment 1. Considering the fact that participants had to solve twice as many problems, we expected that they would be more likely to spontaneously produce the targeted recall mistakes. We thus decided to maintain the minimum sample size of 90 that had been

previously calculated. Participants were students from a second-year university psychology class at the University of Burgundy. They participated in exchange for course credit. A total of 104 students participated in the experiment. Participants all spoke French fluently. One participant was excluded from the analysis due to all his answers being insufficiently detailed to be interpreted. The analyses were conducted on the remaining 103 participants (72 women,  $M_{\text{age}} = 20.43$  years,  $SD = 1.43$ ).

## Materials

The problems used in this experiment were the same as those used in Experiment 1, the difference being that every participant was asked to solve the four problems, instead of two problems being randomly selected among four in the previous experiment.

## Procedure

The experiment was conducted collectively, in a university classroom. Each participant was given a five-page booklet with the following instructions written on the front page:

You will find an arithmetic problem on each page of this booklet. Your task is to solve the problems using as few operations as possible. You can use the “draft” area, but please copy in the “response” area all the operations that you used to come up with the solution. We ask that you take enough time to read and understand each of these problems, as this is not a speed test. Remember that the goal is to solve the problems using as few operations as possible. For every problem, we ask that you write down every operation(s) that you used to come up with the solution, even the simplest one(s) that you can mentally calculate. For instance, the computation “ $15 - 6 - 2 = 7$ ” should not be written as a unique operation, but broken down as “ $15 - 6 = 9$ ” and “ $9 - 2 = 7$ ,” which then count for two operations. *Translated from French*

The four following booklet pages were divided into three parts: the problem statement, the *draft* area, and the *response* area. Problem order and numerical values were controlled across four booklets: half of them introduced the problems in a specific order (Collection 1; Duration 2; Collection 2; Duration 1), and the other half used the reverse order (Duration 1; Collection 2; Duration 2; Collection 1). Thus, half the participants started with a cardinal problem, and the other half started with an ordinal problem. When participants were done solving the problems, their booklets were collected, and new booklets were, then, handed out to them. Here, the instructions read “On the following pages, you will be asked to recall the text of the problems you just solved. Try to write down the problem statements as faithfully as possible, from memory” (translated from French). Then, on each following page, the recall of a specific problem statement was cued using a sentence describing the theme of the problem. For example, “Try to write down, from memory, the text of the first problem you had to solve, about Paul and Jolene’s marbles” (translated from French). The problems had to be recalled in the order in which they originally appeared, to avoid any recency effect.

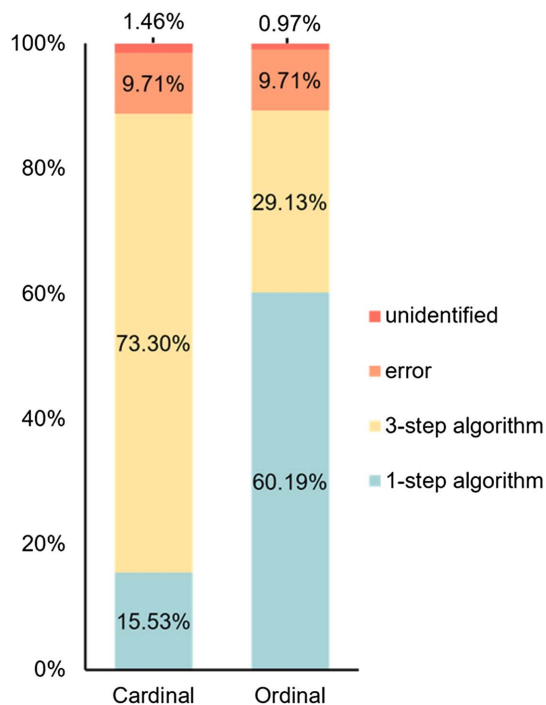
## Results

As in Experiment 1, we studied participants’ responses both in the solving task and in the problem recall task, to investigate the differences between their representations of cardinal and ordinal problems. In 98.79% of the cases, the strategies used by the

participants to solve the problems could easily be inferred from their report of the operations they used to solve the problems. Their responses were either classified as “one-step strategy” (successful use of the shortest strategy), “three-step strategy” (successful use of the longest strategy), or “error” (failure to use any relevant strategy to solve the problem). The 1.21% of answers that were not detailed enough to be analyzed were classified as “unidentified.” The distribution of the participants’ solving strategies depending on the ordinal versus cardinal nature of the problems is described in Figure 4. As in Experiment 1, a generalized linear mixed model with a binomial distribution was used on the 412 recorded trials to evaluate how likely participants were to use the one-step strategy on cardinal and on ordinal problems. The dependent variable was the successful use of the one-step strategy, and the semantic nature of the problems (cardinal vs. ordinal) was included as a fixed effect. We accounted for each participant solving each type of problem by including a random effect for participants. To ascertain the most fitting model, we compared one with only a random intercept for participants against another with by-participant random slopes for the effect of semantics. The latter model demonstrated a significantly better fit, as indicated by a lower Akaike information criterion (372.28 vs. 382.98) and Bayesian information criterion (BIC; (392.38 vs. 395.04), along with a reduced deviance (362.28 vs. 376.98);  $\chi^2(2) = 14.70$ ,  $p < .001$ . Due to concerns of model singularity, we refrained from including random effects for variations between problem statements. The selected model converged successfully and accounted for 97.2% of the total variance (conditional  $R^2$ ). The significance of the fixed effect of semantics (Estimate = 12.17,  $SE = 3.45$ ,  $z = 3.53$ ,

**Figure 4**

*Strategy Distribution in Experiment 2 Depending on the Quantities Used in the Problems*



*Note.* See the online article for the color version of this figure.

$p < .001$ ) indicates that, as predicted in Hypothesis 1, participants were considerably more likely to discover the one-step strategy in ordinal problems (60.19%) than they were in cardinal problems (15.53%). The random intercept for participants indicated substantial variability (Variance = 148.1,  $SD = 12.17$ ), as did the random slope for the effect of semantics across participants (Variance = 115.0,  $SD = 10.73$ ).

Table 4 presents the distribution of erroneous recalls of a whole-to-whole difference instead of the part-to-part difference on cardinal and ordinal problems. As in Experiment 1, erroneously recalling a whole-to-whole difference was interpreted as a sign that participants had made the inference that the difference between the parts was equal to the difference between the wholes, which was thought to be favored within an ordinal encoding but not within a cardinal encoding. Participants recalled this inference instead of the proper problem phrasing in 9.71% of the ordinal problems, whereas they never once made this recall mistake on the cardinal problems. Because no participant recalled a whole-to-whole difference on either of the two cardinal problems, we used an exact McNemar test to analyze the difference between the recall mistakes on cardinal and on ordinal problems. The test indicated that the semantic nature of the problems (ordinal vs. cardinal) had a significant impact on the rate of whole-to-whole sentences being erroneously recalled by the participants,  $\chi^2(1, N = 206) = 151.44, p < .001$ , thus supporting Hypothesis 2. Interestingly, again the number of recall mistakes was higher on the Duration 2 problem ( $N = 13$ ) than on the Duration 1 problem ( $N = 7$ ), although the low number of mistakes overall prevented an in-depth analysis of this potential difference.

Finally, we studied which of the strategies used in the problem-solving task were the most likely to lead participants to erroneously recall a whole-to-whole sentence. Since no participant recalled a whole-to-whole sentence on any cardinal problems, all the cases of erroneous inferences came from ordinal problems. Table 5 details the strategies used prior to misremembering the problems. Interestingly, no use of the three-step strategy was ever followed by a whole-to-whole recall mistake, in line with Hypothesis 3. Instead, 90% of the recall mistakes came from the use of the one-step strategy, one case was associated with failure to solve the problem in the first place, and one was made by a participant whose answer on the related solving task could not be interpreted for lack of detail.

## Discussion

This second experiment replicated and extended the findings of Experiment 1. In the solving task, as expected, participants remained more prone to use the one-step strategy on ordinal than on cardinal problems. In the recall task, the difference observed in Experiment 1 was also replicated. In fact, the distinction between cardinal and

**Table 5**

*Strategies Leading to the Erroneous Recall of Whole-to-Whole Sentences on Ordinal Problems*

| Solving strategy    | No whole-to-whole difference recalled | Whole-to-whole difference erroneously recalled |
|---------------------|---------------------------------------|--|
| One-step strategy   | 106                                   | 18   |
| Three-step strategy | 60                                    | 0  |
| Error               | 1                                     | 1  |
| Unidentified answer | 19                                    | 1  |

ordinal problems was so decisive that out of the 206 attempts to recall a cardinal problem, not a single one led to the erroneous recall of a sentence describing a whole-to-whole difference. The recall of ordinal problems, on the other hand, included this recall mistake about once every 10 trials. This suggests that the differences in the representations encoded by the participants significantly influenced their answers on both tasks in this experiment as well. These results are in line with SECO's predictions regarding the influence of world semantics on the initial encoding of arithmetic word problems. Participants' encoding of the problems included the whole-to-whole inference only on collection problems, which led to specific errors in the recall of these problems. Finally, the analysis of the solving strategies precluding the erroneous recall of a whole-to-whole difference supported the hypothesis of different representations being encoded. Indeed, 90% of the recall mistakes followed the successful use of the shortest strategy, whereas strictly none of the erroneous recalls were preceded by the use of a three-step strategy. Thus, the results supported SECO's depiction of a strong link between participants' representations of the numerical situations and their use of a specific solving strategy. In fact, the correlation between their choice of a solving strategy and their propensity to make a recall mistake shows that the recall mistakes cannot be attributed to a difference in wording between cardinal and ordinal problems. Indeed, ordinal problems were solved using the three-step strategy in 29.1% of the cases, but none of the recall mistakes followed the use of this strategy.

Experiments 1 and 2 relied on participants' tendency to spontaneously make specific mistakes in their attempts to retell the problems. This experimental paradigm makes it possible to investigate the representations that participants constructed, memorized, and freely recalled. There was however one aspect on which Experiment 2 fell short: By doubling the number of problems, we were hoping to increase the task difficulty and thus increase the number of recall mistakes. Yet, participants' rate of targeted recall mistakes overall was lower in this task (4.85%), compared to Experiment 1 (8.27% overall). A possible explanation for this is the fact that the setting in Experiment 2 (a university classroom) was arguably more conducive to concentration than the online survey used in Experiment 1, thus leading to higher performances overall. Another possible factor accounting for this lower rate of recall mistakes is linked to differences between the two populations. Participants in Experiment 2 were university students, likely to have a reasonable level of mathematical proficiency, typical of the general college population. In contrast, participants from Experiment 1 were recruited online and were not screened for mathematics competence in any way. In line with these

**Table 4**

*Distribution of Participants' Erroneous Whole-to-Whole Recalls on Cardinal and Ordinal Problems*

| Problem type      | No mention of a whole-to-whole difference | Erroneous recollection of a whole-to-whole difference |
|-------------------|---|---|
| Cardinal problems | 206 (100%)                                | 0 (0%)  |
| Ordinal problems  | 186 (90.29%)                              | 20 (9.71%)  |

explanations, participants' performance in the solving task did improve in Experiment 2 compared to Experiment 1. In the first experiment, the rate of error and/or unidentified answers in the solving task was 18.11% for cardinal problems and 19.52% for ordinal problems. In this experiment, there were fewer mistakes in the solving task, with only 11.17% of erroneous/unidentified answers on cardinal problems and 10.68% on ordinal problems.

In order to reach a finer understanding of how often participants construct a representation including the whole-to-whole inference, without relying on their ability to spontaneously recall each sentence of the problems, we designed a third experiment involving a sentence-recognition task. The idea was that some participants may have drawn the whole-to-whole inference from their ordinal encodings, without necessarily including it in their retelling of the problems, since the whole-to-whole and part-to-part differences are not mutually exclusive. By asking participants to identify experimenter-induced changes in the fifth sentence of the problem statements, we hoped to get a finer measure of how frequently they included the whole-to-whole inference into their representations.

### Experiment 3

In this third experiment, we used a sentence-recognition paradigm to directly investigate whether participants' representations included the whole-to-whole difference. Instead of recording participants' mistakes in a recall task, we gave them target sentences presenting the difference in the problems either as a part-to-part relation (original sentence) or as a whole-to-whole one (modified sentence). The detection of experimenter-induced changes had notably been used in the analogy literature as a means of evaluating specific aspects of participants' representations that may not necessarily appear within a free recall task (Popov et al., 2017; Silliman & Kurtz, 2019; Vendetti et al., 2014). We thus intended to assess the validity of the three hypotheses formulated in Experiments 1 and 2 using this new experimental paradigm. For this last experiment, we recruited English-speaking participants both for practical reasons (online recruitment through Mechanical Turk) and to strengthen the cross-linguistic robustness of the effects described in the whole study.

## Method

### Participants

Considering that Experiment 3 involved a sentence-recognition task instead of the recollection task used in Experiments 1 and 2, we based our sample size estimation on a previous study by Thevenot (2010) using a similar paradigm. With the BUCSS R package (Anderson & Kelley, 2018), we estimated a minimum sample size of 60 for the sentence-recognition task, for a high level of targeted statistical power (0.9). We recruited 80 participants residing in the United States through the Amazon Mechanical Turk website. All the participants were holders of the "Master Worker" MTurk qualification at the time of the experiment, indicating a high-reliability rating by the platform. We removed from the analyses the 10 participants who stated that they were not native English speakers to avoid potential confounds in problem interpretation. Additionally, we removed three participants who did not manage to solve more than 10% of the problems, thus showing poor attention

during the task. The analyses were conducted on the remaining 67 participants (26 women,  $M_{\text{age}} = 39.18$ ,  $SD = 10.85$ ).

### Materials

In this third experiment, problems were written in English. Each participant was presented with 18 problems: the 12 problems created in Gros et al. (2021), as well as six new problems added to create a pool of 18 problems to choose from. In order to limit the repetitiveness of the task for the participants, we varied the quantities used in the different problems: the pool of ordinal problems was composed of three duration problems, three height problems, and three elevator problems, whereas the pool of cardinal problems was composed of three collection problems, three price problems, and three weight problems. Each of these quantities had been used in previous experiments investigating the role of the cardinal versus ordinal dimension (Gros et al., 2021). We used a within-subject design to allow for within-subject comparisons between performance on cardinal and on ordinal problems.

In the recognition task, two types of problem statements were presented to the participants: problem statements identical to the original ones (one third of the trials) and problem statements in which one sentence had been modified to present the difference as a whole-to-whole difference instead of the part-to-part difference in the original wording (two thirds of the trials). Examples of such modifications for the Duration 1 and Collection 1 problems are presented in Table 6. Both versions for all 18 problem statements can be found in Tables A2 and A3.

### Procedure

This experiment was conducted online using the Qualtrics platform for online experiments. On the first page, the instructions read:

On the following pages, you will be presented with a series of short math problems. Your task is to solve the problems using as few operations as possible. We ask that you take enough time to read and understand each of these problems, as this is not a speed test. Remember that the goal is to solve the problems using as few operations as possible. For every problem, we ask you to type down every operation(s) that you used to come up with the solution, even the simplest one that you can calculate mentally. For instance, the computation " $15 - 6 - 2 = 7$ " should not be written as a unique operation, but broken down as " $15 - 6 = 9$ " and " $9 - 2 = 7$ ," which then count for two operations.

A different problem statement was displayed on each of the 18 following pages. We used nine cardinal problems and nine ordinal problems (see Tables A2 and A3, column "Original problem statement"). A random problem order was assigned to each participant by the Qualtrics platform.

When participants had answered every problem, they were presented with a short nonmathematical distractor task designed to increase the rate of mistakes in the following recognition task by spacing out the solving and the recognition tasks. The distractor task consisted of three short, nonnumerical situations in which participants had to select an explanation for a natural phenomenon among three different interpretations. When participants had chosen an explanation for the three situations, they were then presented with the unexpected recognition task. The following instructions were displayed:



**Table 6***Illustration of the Modifications Introduced in the Problem Statements for the Recognition Task of Experiment 3*

| Problem identifier | Original problem statement  | Modified problem statement  |
|--------------------|---|---|
| Duration 1         | Sofia travelled 5 hours.<br>Her trip started during the day.<br>Sofia arrived at 11 h.<br>Fred left at the same time as Sofia.<br>Fred's trip lasted 2 hours less than Sofia's.<br>What time was it when Fred arrived?                                      | Sofia travelled 5 hours.<br>Her trip started during the day.<br>Sofia arrived at 11 h.<br>Fred left at the same time as Sofia.<br>Fred arrived 2 hours before Sofia.<br>What time was it when Fred arrived?                           |
| Collection 1       | Paul has 7 red marbles.<br>He also has blue marbles.<br>In total, Paul has 13 marbles.<br>Jolene has as many blue marbles as Paul, and some green marbles.<br>She has 2 green marbles less than Paul has red marbles.<br>How many marbles does Jolene have? | Paul has 7 red marbles.<br>He also has blue marbles.<br>In total, Paul has 13 marbles.<br>Jolene has as many blue marbles as Paul, and some green marbles.<br>She has 2 marbles less than Paul.<br>How many marbles does Jolene have? |

*Note.* On all 36 problems, changes were systematically introduced in the fifth sentence.

In the next part of this experiment, you will be presented with a series of problem statements. Some of these problems will be strictly identical to the ones you solved in the first part of the experiment, and some will be slightly different. For each problem, a sentence will be highlighted in red. Your task will be to decide, for each problem, whether the sentence highlighted in red is the same as before or whether it has been modified. The part of the text that is not highlighted in red will be no different in either case. Please read the problem statements entirely and take the time to understand them, as this is not a speed test.

Participants were then presented with a series of 18 problem statements to evaluate. The fifth sentence was systematically highlighted in red, and participants had to answer the question "Is the sentence highlighted in red the same as before?" Two thirds of the problems were presented in their modified version (right column of Tables A2 and A3); these 12 target problems were the focus of our analyses. To partially fulfill subjects' natural expectations regarding the distribution of "original"/"modified" answers, we also introduced six unmodified problems, acting as distractors. They were identical to the problem statements that had been presented in the solving task (left column of Tables A2 and A3). Each participant was assigned to one of three possible combinations of modified/unmodified problems, where either all problems with identifiers ending in "1," all problems ending in "2," or all problems ending in "3," remained unmodified. The resulting set of 12 modified and six unmodified problems was then randomized by the platform for each participant.

## Results

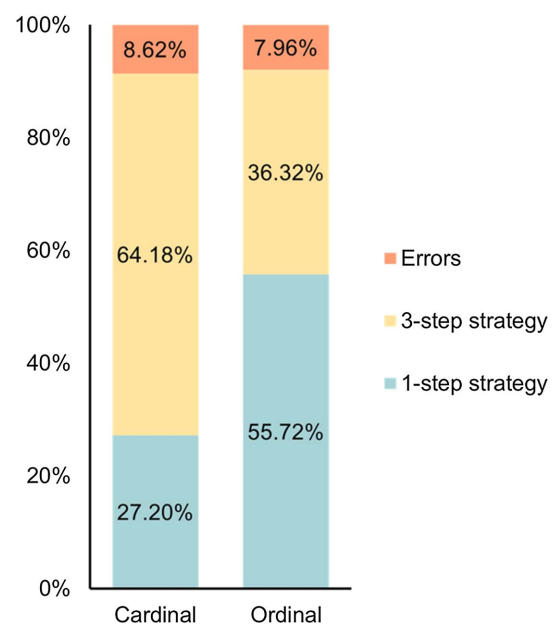
There was no case of "unidentified" answers to report in this experiment, so each response was either classified as "one-step strategy" (successful use of the shortest strategy), "three-step strategy" (successful use of the longest strategy), or "error" (failure to use any relevant strategy to solve the problem). The distribution of the participants' solving strategies depending on the ordinal versus cardinal nature of the problems is described in Figure 5.

As in Experiments 1 and 2, we used a generalized linear mixed model with a binomial distribution to evaluate how likely participants were to use the one-step strategy on cardinal and on ordinal problems, on the 1,206 recorded trials. We selected the successful use of the one-step strategy as the dependent variable and

the semantic nature of the problems as a fixed effect. We included random effects at the participant and problem levels. As in Experiment 2, we compared the fit of a model including random intercepts for the participants and one including random slopes for the effect of semantics across participants. The model with random slopes showed significantly better fit, as evidenced by a considerable reduction in both Akaike information criterion (990.94 vs. 1031.24) and BIC (1021.5 vs. 1051.6), alongside a marked decrease in deviance (978.94 vs. 1023.24);  $\chi^2(2) = 44.29, p < .001$ . The selected model thus included a random intercept for each problem (Variance = 0.04,  $SD = 0.19$ ) as well as random slopes for the effect of semantics across subjects (Variance = 5.58,  $SD = 2.36$ ).

**Figure 5**

*Strategy Distribution in Experiment 3 Depending on the Quantities Used in the Problems*



*Note.* See the online article for the color version of this figure.

Analysis of the fixed effect showed that, as in the previous experiments and in accordance with Hypothesis 1, the problems were more likely to be solved using the one-step strategy when they featured ordinal quantities (55.72%) than when they featured cardinal quantities (27.20%); Estimate =  $-3.24$ ,  $SE = 0.54$ ,  $z = 6.01$ ,  $p < .001$ ,  $R^2_{GLMM(c)} = .818$ . When analyzing the data at the problem level, interestingly, the percentage of use of the one-step strategy was systematically higher in ordinal problems (between 46.6% and 66.7%) than in cardinal problems (from 16.7% to 31.8%). This, along with the limited variance of the random effect for problem in the mixed model, suggested a consistent effect of semantics across problem statements (see Figure B1 for the complete distribution of solving strategies at the problem level).

Regarding the recognition task, we studied how likely participants were to falsely recognize modified problems, depending on the cardinal versus ordinal nature of the problems, as well as on the solving strategies they used in the solving task. Since it has been shown that participants who fail to solve a problem tend to make a random choice in a subsequent recognition task (Hegarty et al., 1995; Thevenot, 2010), we removed the recognition data corresponding to the 8.29% of trials where participants did not successfully solve the problem, resulting in 1,106 trials in which the problems had been correctly solved with either strategy in the solving task. We then selected the 739 trials where participants were shown a target problem involving a modified sentence and removed the data collected on unmodified distractors, to assess when participants would erroneously recognize modified sentences as original. We used a generalized linear mixed model with a binomial distribution to identify which factors influenced the participants' responses on the modified problems. We used the response to the recognition task as the dependent variable, the semantic nature of the problems and the solving strategy as two fixed effects, and we accounted for each participant solving both types of problems by including random effects at the participant and problem level. As previously, we compared two models: one including random intercepts only and another incorporating random slopes for the effect of semantics across participants. The chi-square test did not significantly favor one model over the other,  $\chi^2(2) = 5.68$ ,  $p = .058$ . While the model with random slopes exhibited a marginally lower Akaike information criterion (780.10 vs. 781.78), suggesting a slightly better fit, this was not reflected in the BIC values, with the simpler model showing a lower score (804.81 vs. 812.34). Given the stronger penalty for complexity in BIC and in line with the principle of parsimony, we chose the simpler model with only random intercepts for our final analysis. This model thus included a random intercept for participants (Variance = 1.59,  $SD = 1.26$ ), as well as a random intercept accounting for variations between problem statements (Variance = 0.23,  $SD = 0.48$ ).

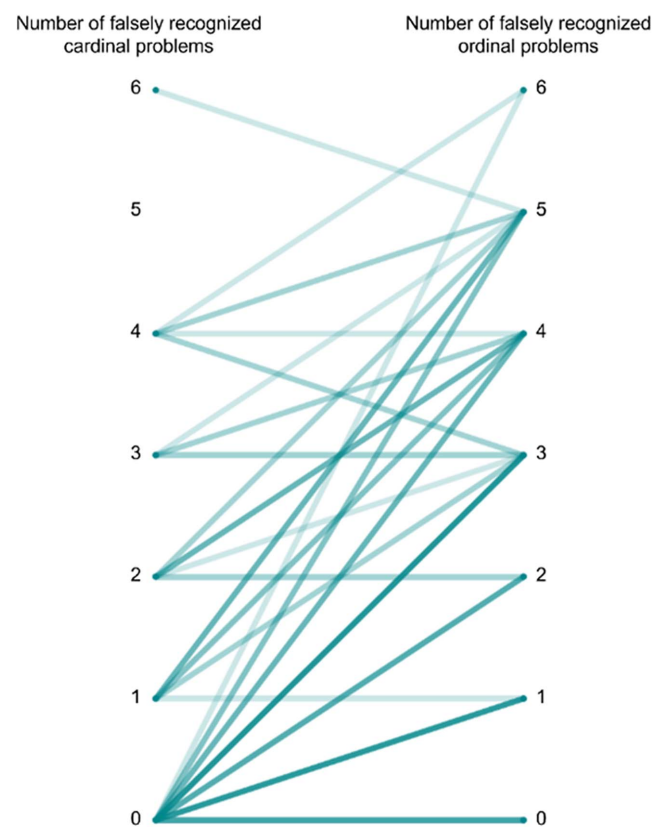
The model successfully converged with a total explanatory power of 48.6% (conditional  $R^2$ ). Results showed that, as hypothesized (Hypothesis 2), participants were more likely to incorrectly recognize the modified problems when they were ordinal problems (52.14% of false recognition) than when they were cardinal problems (16.99% of false recognition); Estimate =  $2.07$ ,  $SE = 0.32$ ,  $z = 6.54$ ,  $p < .001$ . At the problem level, despite some degree of variations between problem statements, the rate of false recognition of altered sentences was systematically higher for any ordinal problem (from 47.6% to 66.7%) as compared to any cardinal

problem (from 16.7% to 31.8%; see Figure B2 for a description of the percentage of false recognition on each problem).

To further explore the part of the variance explained by between-participants differences, we also looked at interindividual variations in the percentage of false recognitions depending on the type of problems. Each participant had been presented six modified cardinal problems and six modified ordinal problems; we looked at the number of modified problems that each participant mistakenly accepted (see Figure 6). Interestingly, most participants (77.61%) misremembered more ordinal problems than cardinal problems, 17.91% of the participants misremembered as many ordinal problems as cardinal problems, and only 4.48% of the participants misremembered more cardinal than ordinal problems. Thus, even at an individual level, only three participants displayed a pattern of misrememberings that went against our hypotheses.

Finally, consistent with our third hypothesis (Hypothesis 3), participants' rate of false recognition was also dependent on which

**Figure 6**  
*Slope Graph Describing the Individual Trajectories in Experiment 3*



*Note.* This graph describes participants' total number of false recognitions in two categories of problems—cardinal (left) and ordinal (right). Each line represents a specific response profile, connecting their performance across the two problem types. The darkness of the line corresponds to the number of participants exhibiting a similar trajectory, with darker lines indicating a more common profile among participants. Most slopes indicate an increase between cardinal problems and ordinal problems, with only three participants presenting a decreasing pattern indicating a higher rate of false recognition of cardinal problems. See the online article for the color version of this figure.

strategy they used to solve the problems in the first place: Participants who successfully solved a problem were more likely to incorrectly recognize a modified problem if they had solved the original problem with the one-step strategy (45.40% of false recognition) than if they had solved it with the three-step strategy (26.39% of false recognition), even after accounting for the effect of the cardinal–ordinal distinction; Estimate = 0.52,  $SE = 0.25$ ,  $z = 2.05$ ,  $p = .041$ . Thus, both the semantic nature of the problems and the ability to use the one-step strategy in the solving task were independently linked to participants' tendency to mistake a modified sentence for an original one. The highest rate of false recognition of altered sentences was observed for ordinal problems solved in one step (56.16%), whereas the lowest error rate was reached on cardinal problems solved in three steps (14.34%; see Figure 7).

## Discussion

This third experiment brought new evidence that the distinction between cardinal and ordinal quantities has a crucial role on the representation of arithmetic word problems, displaying a strong enough influence that it shaped participants' recollection of the described situations, as well as of the strategies they used to solve the problems. Across 18 different contexts, participants' solving strategies were significantly influenced by the quantities involved in the problems, which supports the idea that the representations they constructed were different on cardinal and on ordinal problems. The sentence-recognition task revealed that participants' memory of the different situations differed between cardinal and ordinal problems. Using a sentence-recognition task made it possible to pinpoint specific aspects of their representations following our predictions. Results were aligned with the hypothesis that the encoding of the ordinal problems tended to include the whole-to-whole inference,

whereas the encoding of the cardinal problems did not usually convey this piece of information.

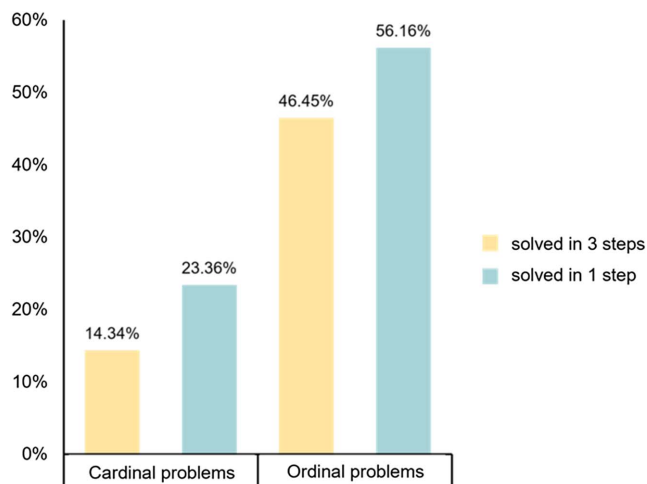
Crucially, the ability to use the one-step strategy in the solving task was also a predictor of participants' rate of false recognition of the modified problems, even after the effect of the cardinal versus ordinal dimension was accounted for. In other words, using the one-step strategy, regardless of the nature of the quantities it involves, was linked to an increase in the probability of falsely recognizing its modified version. The fact that this effect was observed on cardinal and on ordinal problems alike corroborates SECO's prediction (b) regarding the possibility to construct different encodings of the same problem statement. The participants who managed to go beyond their initial encoding of a cardinal problem and use the one-step algorithm to solve it were more likely to falsely recognize the modified sentence on these problems, suggesting that they recoded their initial cardinal representation into a new representation including the targeted inference. In other words, this strategy difference can be traced back to an encoding (or recoding) difference, as evidenced by the false recognition of the whole-to-whole inference.

Of note, the number of whole-to-whole erroneous recognitions was higher in this experiment than the number of whole-to-whole erroneous recalls in Experiments 1 and 2. This suggests that the use of a sentence-recognition task did increase our ability to identify the presence of the whole-to-whole inference within participants' representations, in line with the rationale for this third experiment. Additionally, it should be noted that since participants had to solve 18 problems before being asked to recognize target sentences from the problems, they had to wait longer between the reading of the first problem and the start of the memory task. It also meant that they would be exposed to more potentially distracting materials before needing to recognize the target sentences, which may also account for a portion of this increase in recall mistakes.

## General Discussion

Going back to the three objectives we set for this article, what can we conclude from the collected data? (1) First, we aimed to evaluate two key predictions of the SECO model. The first prediction (a) stated that general, nonmathematical knowledge would interfere with the encoding of arithmetic word problems and lead participants to construct specific mental representations of the problems they encounter, which would in turn influence the solving strategies they use. The data collected in Experiments 1, 2, and 3 supported this hypothesis: while the problems all shared the same mathematical structure, the mention of different types of quantities used in daily-life situations led to different responses. On the problems mentioning ordinal quantities, participants were more likely to erroneously recall or recognize a sentence describing a whole-to-whole difference, despite this sentence being absent from the original problems. This is consistent with the idea that quantities attached to ordinal world semantics lead participants to abstract an ordinal representation that makes it easier to infer that the part-to-part difference is equal to the whole-to-whole difference. On problems mentioning cardinal quantities, however, the error was noticeably rarer, since making this inference from a cardinal perspective is considerably less trivial. Additionally, in all three experiments, participants were significantly more likely to use the

**Figure 7**  
*Rate of False Recognition of Altered Sentences, Depending on the Quantities Used in the Problems and on the Strategies Used to Solve Them*



Note. See the online article for the color version of this figure.

one-step algorithm on ordinal problems than on cardinal problems, despite being instructed to use the shortest possible solving strategies.<sup>1</sup> This also supports SECO's claim that different world semantics result in different encodings being constructed and subsequently lead to different solving strategies being used. All the problems could have been solved in one step, yet the vast majority of participants used three steps to solve the problems with cardinal semantics. The fact that problems mentioning elevators and problems mentioning durations, despite their many surface differences, would lead to the same solving strategies as well as to the same false memories is something that the situation model framework would not have predicted. Similarly, the explanation as to why marble problems and duration problems led to different responses in all three experiments falls outside the scope of the schema framework.

The second prediction (b) brought by SECO regarded participants' ability to recode their initial encoding of the situation and construct a new representation in an effort to use a solving strategy that was not semantically congruent with their initial representation. Data analysis in Experiment 3 revealed that, both on cardinal problems and on ordinal problems, the participants who managed to solve a problem in one step were also more likely to falsely recognize a sentence from this problem describing the whole-to-whole difference. Thus, regardless of the ordinal/cardinal semantics attached to the problem statements, participants' use of the one-step strategy was associated with the nature of the representation that guided their replies in the sentence-recognition task. This suggests that the participants who used the one-step strategy on cardinal problems did so by going beyond a cardinal encoding of the situation and constructing a representation that included the inference that underlies the use of the one-step strategy. This supports SECO's claim that the use of a specific solving strategy depends on the structure of the representations that were constructed, either initially or after recoding the initial representation.

Our next objective (2) with this study regarded the contribution to the growing line of evidence that the perception of cardinality and ordinality has a deep impact on mathematical reasoning, even among adults. By focusing on problems allowing two strategies, each compatible either with a cardinal representation or with an ordinal representation, we aimed to further advance the understanding of this dimension. In all three experiments, across 18 different problems and two different languages, we were able to extend the findings of Gros et al. (2021) regarding the deep influence that the perception of cardinality and ordinality has on adults' performance. The adults in Experiments 1, 2, and 3 only managed to use the one-step strategy on, respectively, 12.60%, 15.50%, and 24.90% of the cardinal problems, whereas they used it on the majority of ordinal problems (51.22%, 60.19%, and 53.80%, respectively). Additionally, the encoding differences led participants to evoke false memories of what they had read in the recall tasks. They tended to include specific inferences regarding the mathematical structure of the problems on ordinal problems but not on cardinal problems, thus showing the significant impact that cardinality and ordinality can have on human reasoning. We believe that this new evidence constitutes an important step forward in understanding and characterizing the extent to which this dimension plays a central role in adults' numerical cognition, beyond the initial learning of numbers (e.g., Colomé & Noël, 2012).

Finally, our last objective (3) was to investigate whether false memories can be used as a means to distinguish between different

mathematical representations of a given problem. Since direct inspection of mental constructs is hardly feasible, numerous indirect routes have been proposed by cognitive scientists aiming to scrutinize the representations underlying human thought processes (Pearson & Kosslyn, 2015). In the field of mathematical cognition, insights have come from behavioral measures such as response times (e.g., Dehaene et al., 1993), gestures (e.g., Brooks et al., 2018), drawings (Rellensmann et al., 2017), eye movements (e.g., Strohmaier et al., 2020), or metaphors (e.g., Lakoff & Núñez, 2000), but also from physiological measures such as fMRI activation patterns (e.g., Cohen Kadosh et al., 2007), event-related potentials (e.g., Bagnoud et al., 2018), or pupil dilation (e.g., Salvaggio et al., 2022). However, false memories have received less attention than they deserve in the literature on mathematical cognition. Building upon seminal studies on problem recall (Verschaffel's, 1994) and problem recognition (Thevenot, 2010), we wished to go one step further and explore how problem recall and sentence recognition could help identify specific differences between the representations of isomorphic problems.

We believe the results brought by the three unexpected memory tasks (through either problem recall or sentence recognition) provide crucial insights into the nature of the constructed representations. We have known ever since Loftus' work on the creation of false memories in long-term memory that recall and recognition can be tempered with by leading participants to represent a situation they have never actually experienced (Loftus, 1996; Loftus & Pickrell, 1995). More recently, it has been shown that false memories can also happen at short term when working memory maintenance is impaired by a competing task (Abadie & Camos, 2019), as was the case in our experiment. Here, although we did not plant entirely false memories in the participants' minds, we led them to misremember mathematical information about the scenes described in the problems, by eliciting one of two contrasting encodings of mathematically equivalent situations in working memory. Differences appeared between problems with cardinal quantities and those with ordinal quantities, as well as between identical problems that participants solved with different strategies. By targeting an inference compatible with one of the two predicted representations, our measure makes it possible to probe participants' mental representations without relying on their ability to solve the problems and accurately write down the operations they used to do so. When combined with other behavioral and physiological indicators, these tasks play a crucial role in drawing a full picture of the semantic determinants of mathematical word problem solving.

Interestingly, by showing that new, inferred, mathematical relations that were not explicitly presented in the problem statement are nevertheless encoded and memorized when participants attempt to solve arithmetic word problems, these three experiments also shed light on a current question in the literature on analogical reasoning. Indeed, a line of research has sought to determine which aspects of situations are encoded and later recalled, insisting on the difference between surface features (the objects' attributes that can

<sup>1</sup> In consideration of the potential differences in working memory demands across ordinal and cardinal problems, we conducted an additional experiment controlling the influence of the mention of entities differing by their physical properties in some of the cardinal problems. Given the detailed nature of this experiment and in order to maintain the focus of the article, the specifics, including methodology and results, are provided in Appendix C.



vary from one problem to another without affecting the solution path) and structural ones (the causally relevant relations underlying the problem and impacting the achievement of the goal) (Gentner, 1983; Gentner & Maravilla, 2018; Holyoak, 1985, 2012). In our study, participants were more likely to encode the whole-to-whole inference on problems involving ordinal quantities, which supports the view that they did not systematically reach the abstract structure of the described situations (the problems' deep structure). This observation is consistent with previous studies showing that participants generally have a hard time extracting the abstract schema underlying a problem statement (Gick & Holyoak, 1980; Kubricht et al., 2017; Kurtz & Loewenstein, 2007). However, this does not mean that participants only relied on the surface features of the situations either (cf. Forbus et al., 1995; Gentner et al., 2009), since when dealing with problems involving ordinal quantities, a significant part of the participants did integrate the relational whole-to-whole inference within their encoding, despite it not being explicitly described in the text. The fact that participants neither systematically perceived the problems' deep structure nor remained entirely limited to the surface features in the problem statements supports the view that they encoded an interpreted structure at an intermediate level of abstraction. The SECO model, in line with recent evidence regarding this issue (Raynal et al., 2020), suggests that most participants encode an interpreted structure that relies on world knowledge associated with the surface features of the problem statement to integrate some structural aspects of the problems.

One may wonder why individuals rely so heavily on surface features when engaging in such reasoning, considering the inherent abstraction of mathematics (Davis et al., 2011). The fact that the simple mention of one quantity over another (e.g., hours instead of marbles) would dictate whether participants will be able to infer a specific relational statement between two mathematical entities of problems may seem, at first, like a deep flaw of human reasoning. However, one possible explanation comes from the realization that the surface features of situations are generally correlated with their deeper principles (Bassok et al., 1995; Blessing & Ross, 1996; Gentner & Medina, 1998; Trench & Minervino, 2015). As expressed by Gentner through the *kind world* hypothesis, "what looks like a tiger is very likely to be a tiger, with the relational characteristics of a tiger, such as 'eats other animals'" (Gentner & Maravilla, 2018, p 196). This means that using the superficial aspects of situations to infer their deep structure is, more often than not, a fruitful approach. It is not entirely surprising, then, that humans would rely on contextual clues to comprehend the situations they encounter, even if that exposes them to occasional errors in reasoning. Even experts have been shown to (overly) rely on such content effects, sometimes to the point of failing a task within their own field of expertise (Blessing & Ross, 1996; Gros et al., 2019). In fact, the consistent use of the "expert pathway" described in the SECO model (see Figure 1) by experts is still a subject of debate, as even experienced mathematicians have been observed to rely on world semantics, especially under time constraints (Gros et al., 2019). In a similar vein, Blessing and Ross (1996) have highlighted that a correlation exists between problems' surface and structural features. They pointed out that problems involving a similar content (e.g., mention of people's ages at two different times) tend to require the application of an identical solving principle, for example, the same algebraic equations: " $x = Ay$ ;  $x + B = C(x + B)$ ". The authors found evidence that even experienced participants were influenced

by such correlations, performing better when problem content was typical for the problem's deep structure than when it was not. According to them, this finding illustrates that problem schemata retain superficial aspects that specific instances usually preserve. Thus, cues provided by the surface content of a problem are generally informative with regard to the solving strategies to implement, and following those cues is a reasonable heuristic in most situations, even for experts. However, in the present study, we created situations in which the cues to perceive the cardinality in the problems led to longer solving strategies, while the cues to perceive the ordinality in other problems led to recall mistakes in the following tasks. Those examples demonstrate the limitations of content-based heuristics: When the inferences students make based on their knowledge about the world conflict with the mathematical structure of a situation, strong limitations may hinder their ability to use the best (or any) solving strategy (Gros et al., 2020).

On the other hand, this worldly influence may also account for some of the benefits found when using word problems, as compared with more formal problems. For instance, Koedinger and Nathan (2004) investigated the verbal facilitation hypothesis, according to which presenting formal problems (such as algebra problems) in familiar natural language will help cue students' preexisting knowledge and will foster the use of informal solving strategies. In two experiments, they found that high school students were more successful at solving entry-level algebra problems when those were presented as word problems instead of formal equations. The authors concluded that the presentation of problems, notably their contextual embedding, significantly influences students' reasoning. They suggested beginning problem-solving instruction with story-based problems, capitalizing on students' proficiency in this area. According to them, this approach should be followed by a gradual transition to more abstract word problems that involve equations, eventually leading to instruction in entirely symbolic equations. In a related perspective, a line of works suggested that using "authentic" arithmetic word problems, close to real-life situations, may help increase performance in math education (Palm & Nyström, 2009; Vicente & Machado, 2016; Vicente et al., 2021). However, the effectiveness of using authentic situations may be further optimized by systematically considering the impact of SECO. According to the SECO model, failing to account for this and resorting to incongruent world semantics in a problem statement inspired by real life could be counterproductive and potentially detrimental to students' performance.

Adding to these educational considerations, the present study provides new insights into the difficulties met by solvers in their attempts to see past the superficial dissimilarities between problems and perceive the isomorphism between situations (Duque de Blas et al., 2021; Gros et al., 2015; Gvozdic & Sander, 2018; Scheibling-Sève et al., 2020). Whereas prior research has suggested that transfer crucially depends on the implementation of a uniform relational encoding across isomorphs (Gentner, 2010), our findings indicate that participants tend to focus on different types of relations (i.e., whole-to-whole or part-to-part) according to the nature of the quantities involved in the problem. This incompatibility between the representations encoded from isomorphic problems may provide an account of the systematic failures that have been documented in the transfer literature (Day & Goldstone, 2012) and raises new questions regarding how to help participants recode their initial representations in a way that would promote transfer. Namely, a growing body

of research has investigated promising methods that might improve transfer to some extent, such as inviting participants to compare or categorize isomorphic problems (Gamo et al., 2010; Gvozdic & Sander, 2020; Iacono et al., 2022; Kurtz & Honke, 2020; Scheibling-Sève et al., 2020, 2022) or providing idealized or animated representations of the problems (Goldstone & Sakamoto, 2003; Kubricht et al., 2017; Trench et al., 2017). However, the question of what it takes to systematically see the deep structure of arithmetic word problems regardless of the inferences drawn from the context they are embedded in remains a decisive issue that may benefit from further studies investigating learners' memories.

## References

- Abadie, M., & Camos, V. (2019). False memory at short and long term. *Journal of Experimental Psychology: General*, 148(8), 1312–1334. <https://doi.org/10.1037/xge0000526>
- Anderson, S. F., & Kelley, K. (2018). *BUCSS: Bias and uncertainty corrected sample size* [Computer software and manual] (R package Version 1.2.1). <https://CRAN.R-project.org/package=BUCSS>
- 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. <https://doi.org/10.1177/0956797617723724>
- Baccaglini-Frank, A., Carotenuto, G., & Sinclair, N. (2020). Eliciting preschoolers' number abilities using open, multi-touch environments. *ZDM Mathematics Education*, 52(4), 779–791. <https://doi.org/10.1007/s11858-020-01144-y>
- Bagnoud, J., Burra, N., Castel, C., Oakhill, J., & Thevenot, C. (2018). Arithmetic word problems describing discrete quantities: E.E.G evidence for the construction of a situation model. *Acta Psychologica*, 190, 116–121. <https://doi.org/10.1016/j.actpsy.2018.07.008>
- Bassok, M. (2001). Semantic alignments in mathematical word problems. In D. Centner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 401–433). The MIT Press.
- Bassok, M., Wu, L. L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition*, 23(3), 354–367. <https://doi.org/10.3758/BF03197236>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4* (R package Version 1.1–7).
- Bermejo, V. (1996). Cardinality development and counting. *Developmental Psychology*, 32(2), 263–268. <https://doi.org/10.1037/0012-1649.32.2.263>
- Black, J. B., & Bern, H. (1981). Causal coherence and memory for events in narratives. *Journal of Verbal Learning & Verbal Behavior*, 20(3), 267–275. [https://doi.org/10.1016/S0022-5371\(81\)90417-5](https://doi.org/10.1016/S0022-5371(81)90417-5)
- 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. <https://doi.org/10.1037/0278-7393.22.3.792>
- Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11(2), 177–220. [https://doi.org/10.1016/0010-0285\(79\)90009-4](https://doi.org/10.1016/0010-0285(79)90009-4)
- Bransford, J. D., & Johnson, M. K. (1973). Considerations of some problems of comprehension. In W. Chase (Ed.), *Visual information processing* (pp. 383–438). Academic Press. <https://doi.org/10.1016/B978-0-12-170150-5.50014-7>
- Brooks, N. B., Barner, D., Frank, M., & Goldin-Meadow, S. (2018). The role of gesture in supporting mental representations: The case of mental abacus arithmetic. *Cognitive Science*, 42(2), 554–575. <https://doi.org/10.1111/cogs.12527>
- Cheung, C. N., & Lourenco, S. F. (2019). Does  $1 + 1 = 2nd$ ? The relations between children's understanding of ordinal position and their arithmetic performance. *Journal of Experimental Child Psychology*, 187, Article 104651. <https://doi.org/10.1016/j.jecp.2019.06.004>
- 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. [https://doi.org/10.1207/s15516709cog0502\\_2](https://doi.org/10.1207/s15516709cog0502_2)
- Cohen Kadosh, R., Cohen Kadosh, K., Kaas, A., Henik, A., & Goebel, R. (2007). Notation-dependent and -independent representations of numbers in the parietal lobes. *Neuron*, 53(2), 307–314. <https://doi.org/10.1016/j.neuron.2006.12.025>
- Colomé, A., & 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. <https://doi.org/10.1016/j.jecp.2012.03.005>
- Corbett, A. T., & Doshier, B. A. (1978). Instrument inferences in sentence encoding. *Journal of Verbal Learning & Verbal Behavior*, 17(4), 479–491. [https://doi.org/10.1016/S0022-5371\(78\)90292-X](https://doi.org/10.1016/S0022-5371(78)90292-X)
- Dantzig, T. (1945). *Number the language of science: A critical survey written for the cultured non-mathematician*. The Macmillan Company.
- Daroczy, G., Meurers, D., Heller, J., Wolska, M., & Nürk, H. C. (2020). The interaction of linguistic and arithmetic factors affects adult performance on arithmetic word problems. *Cognitive Processing*, 21(1), 105–125. <https://doi.org/10.1007/s10339-019-00948-5>
- 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, Article 348. <https://doi.org/10.3389/fpsyg.2015.00348>
- Dauben, J. W. (1990). *Georg Cantor: His mathematics and philosophy of the infinite*. Princeton University Press.
- Davis, P., Hersh, R., & Marchisotto, E. A. (2011). *The mathematical experience*. Springer.
- Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, 47(3), 153–176. <https://doi.org/10.1080/00461520.2012.696438>
- Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, 122(3), 371–396. <https://doi.org/10.1037/0096-3445.122.3.371>
- Duque de Blas, G., Gómez-Veiga, I., & García-Madruga, J. A. (2021). Arithmetic word problems revisited: Cognitive processes and academic performance in secondary school. *Education Sciences*, 11(4), Article 155. <https://doi.org/10.3390/educsci11040155>
- Fischer, F. E., & Beckey, R. D. (1990). Beginning kindergarteners' perception of number. *Perceptual and Motor Skills*, 70(2), 419–425. <https://doi.org/10.2466/pms.1990.70.2.419>
- Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19(2), 141–205. [https://doi.org/10.1207/s15516709cog1902\\_1](https://doi.org/10.1207/s15516709cog1902_1)
- Frege, G. (1980). *The foundations of arithmetic* (Vol. 1884; J. Austin, Trans.). Northwestern University Press.
- 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. <https://doi.org/10.1016/j.learninstruc.2009.04.001>
- Geary, D. C., & vanMarle, K. (2018). Growth of symbolic number knowledge accelerates after children understand cardinality. *Cognition*, 177, 69–78. <https://doi.org/10.1016/j.cognition.2018.04.002>
- Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Harvard University Press. <https://doi.org/10.4159/9780674037533>
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170. [https://doi.org/10.1016/S0364-0213\(83\)80009-3](https://doi.org/10.1016/S0364-0213(83)80009-3)
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5), 752–775. <https://doi.org/10.1111/j.1551-6709.2010.01114.x>
- Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. D. (2009). Reviving inert knowledge: Analogical abstraction supports relational

- retrieval of past events. *Cognitive Science*, 33(8), 1343–1382. <https://doi.org/10.1111/j.1551-6709.2009.01070.x>
- Gentner, D., & Maravilla, F. (2018). Analogical reasoning. In L. J. Ball & V. A. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 186–203). Psychology Press.
- Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, 65(2–3), 263–297. [https://doi.org/10.1016/S0010-0277\(98\)00002-X](https://doi.org/10.1016/S0010-0277(98)00002-X)
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12(3), 306–355. [https://doi.org/10.1016/0010-0285\(80\)90013-4](https://doi.org/10.1016/0010-0285(80)90013-4)
- Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology*, 46(4), 414–466. [https://doi.org/10.1016/S0010-0285\(02\)00519-4](https://doi.org/10.1016/S0010-0285(02)00519-4)
- 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*, 26(5), 1738–1746. <https://doi.org/10.3758/s13423-019-01628-3>
- 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). Cognitive Science Society.
- Gros, H., Thibaut, J. P., & Sander, E. (2020). Semantic congruence in arithmetic: A new conceptual model for word problem solving. *Educational Psychologist*, 55(2), 69–87. <https://doi.org/10.1080/00461520.2019.1691004>
- Gros, H., Thibaut, J. P., & Sander, E. (2021). What we count dictates how we count: A tale of two encodings. *Cognition*, 212, Article 104665. <https://doi.org/10.1016/j.cognition.2021.104665>
- Gros, H., Thibaut, J.-P., & Sander, E. (2024a). Uncovering the interplay between drawings, mental representations, and arithmetic problem-solving strategies in children and adults. *Memory & Cognition*. Advance online publication. <https://doi.org/10.3758/s13421-024-01523-w>
- Gros, H., Thibaut, J.-P., & Sander, E. (2024b). *Revealing mental representations of arithmetic word problems through false memories: New insights into semantic congruence* [Dataset]. [https://osf.io/5nqev/?view\\_only=6dcf3a21a2c840c6a16dce2bbf419762](https://osf.io/5nqev/?view_only=6dcf3a21a2c840c6a16dce2bbf419762)
- Gvozdic, K., & Sander, E. (2018). When intuitive conceptions overshadow pedagogical content knowledge: Teachers' conceptions of students' arithmetic word problem solving strategies. *Educational Studies in Mathematics*, 98(2), 157–175. <https://doi.org/10.1007/s10649-018-9806-7>
- Gvozdic, K., & Sander, E. (2020). Learning to be an opportunistic word problem solver: Going beyond informal solving strategies. *ZDM Mathematics Education*, 52(1), 111–123. <https://doi.org/10.1007/s11858-019-01114-z>
- 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. <https://doi.org/10.1037/0022-0663.87.1.18>
- Holyoak, K. J. (1985). The pragmatics of analogical transfer. *Psychology of Learning and Motivation*, 19, 59–87. [https://doi.org/10.1016/S0079-7421\(08\)60524-1](https://doi.org/10.1016/S0079-7421(08)60524-1)
- Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199734689.013.0013>
- Hund, A. M., Kannass, K. N., Bove, R., Fairweather, L., Maydew, M., & Monla, A. (2021). Young children's understanding of ordinal and spatial labels. *Cognitive Development*, 58, Article 101041. <https://doi.org/10.1016/j.cogdev.2021.101041>
- Iacono, E., Gros, H., & Clément, E. (2022). *Training flexible categorization to improve arithmetic problem solving: A school-based intervention with 5th graders* [Conference session]. Proceedings of the 44th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Toronto, Canada.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.
- Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43), 18243–18250. <https://doi.org/10.1073/pnas.1012933107>
- Kintsch, W., & Bates, E. (1977). Recognition memory for statements from a classroom lecture. *Journal of Experimental Psychology: Human Learning and Memory*, 3(2), 150–159. <https://doi.org/10.1037/0278-7393.3.2.150>
- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92(1), 109–129. <https://doi.org/10.1037/0033-295X.92.1.109>
- Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394. <https://doi.org/10.1037/0033-295X.85.5.363>
- Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, 29(2), 133–159. [https://doi.org/10.1016/0749-596X\(90\)90069-C](https://doi.org/10.1016/0749-596X(90)90069-C)
- Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, 13(2), 129–164. [https://doi.org/10.1207/s15327809jls1302\\_1](https://doi.org/10.1207/s15327809jls1302_1)
- Kubricht, J. R., Lu, H., & Holyoak, K. J. (2017). Individual differences in spontaneous analogical transfer. *Memory & Cognition*, 45(4), 576–588. <https://doi.org/10.3758/s13421-016-0687-7>
- Kurtz, K. J., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. *Memory & Cognition*, 35(2), 334–341. <https://doi.org/10.3758/BF03193454>
- Kurtz, K. J., & Honke, G. (2020). Sorting out the problem of inert knowledge: Category construction to promote spontaneous transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(5), 803–821. <https://doi.org/10.1037/xlm0000750>
- Lakoff, G., & Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic Books.
- Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105(2), 395–438. <https://doi.org/10.1016/j.cognition.2006.10.005>
- Loftus, E. F. (1996). Memory distortion and false memory creation. *The Bulletin of the American Academy of Psychiatry and the Law*, 24(3), 281–295.
- Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, 25(12), 720–725. <https://doi.org/10.3928/0048-5713-19951201-07>
- Mani, K., & Johnson-Laird, P. N. (1982). The mental representation of spatial descriptions. *Memory & Cognition*, 10(2), 181–187. <https://doi.org/10.3758/BF03209220>
- Meyer, C., Barbiers, S., Weerman, F., Jennifer, S., & Deb, W. (2016). Order and ordinality: The acquisition of cardinals and ordinals in Dutch. In J. Scott, & D. Waughtal (Eds.), *Proceedings of the 40th annual Boston University conference on language development* (pp. 253–266). Cascadia Press.
- Miller, K., Major, S. M., Shu, H., & Zhang, H. (2000). Ordinal knowledge: Number names and number concepts in Chinese and English. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, 54(2), 129–140. <https://doi.org/10.1037/h0087335>
- Miller, S. E., Marcovitch, S., Boseovski, J. J., & Lewkowicz, D. J. (2015). Young children's ability to use ordinal labels in a spatial search task. *Merrill-Palmer Quarterly*, 61(3), 345–361. <https://doi.org/10.13110/merrillpalmer.1982.61.3.0345>



- Noordman, L. G., & Vonk, W. (2015). Inferences in discourse, psychology of. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 12, pp. 37–44). Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.57012-3>
- Palm, T., & Nyström, P. (2009). Gender aspects of sense making in word problem solving. *Journal of Mathematical Modelling and Application*, 1(1), 59–76.
- Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33), 10089–10092. <https://doi.org/10.1073/pnas.1504933112>
- Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, 146(5), 722–745. <https://doi.org/10.1037/xge0000305>
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Raynal, L., Clément, E., & Sander, E. (2020). Are superficially dissimilar analogs better retrieved than superficially similar disanalogues? *Acta Psychologica*, 203, Article 102989. <https://doi.org/10.1016/j.actpsy.2019.102989>
- Rellensmann, J., Schukajlow, S., & Leopold, C. (2017). Make a drawing. Effects of strategic knowledge, drawing accuracy, and type of drawing on students' mathematical modelling performance. *Educational Studies in Mathematics*, 95(1), 53–78. <https://doi.org/10.1007/s10649-016-9736-1>
- Reusser, K. (1990). From text to situation to equation: Cognitive simulation of understanding and solving mathematical word problems. In H. Mandl, E. De Corte, N. Bennet, & H. F. Friedrich (Eds.), *Learning and instruction, European research in an international context* (Vol. II, pp. 477–498). Pergamon Press.
- Russell, B. (1919). *Introduction to mathematical philosophy*. Allen & Unwin.
- Salvaggio, S., Andres, M., Zénon, A., & Masson, N. (2022). Pupil size variations reveal covert shifts of attention induced by numbers. *Psychonomic Bulletin & Review*, 29(5), 1844–1853. <https://doi.org/10.3758/s13423-022-02094-0>
- Sarnecka, B. W., & Lee, M. D. (2009). Levels of number knowledge during early childhood. *Journal of Experimental Child Psychology*, 103(3), 325–337. <https://doi.org/10.1016/j.jecp.2009.02.007>
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans and understanding: An inquiry into human knowledge structures*. Hillsdale.
- Scheibling-Sève, C., Gvozdic, K., Pasquinelli, E., & Sander, E. (2022). Enhancing cognitive flexibility through a training based on multiple categorization: Developing proportional reasoning in primary school. *Journal of Numerical Cognition*, 8(3), 443–472. <https://doi.org/10.5964/jnc.7661>
- Scheibling-Sève, C., Pasquinelli, E., & Sander, E. (2020). Assessing conceptual knowledge through solving arithmetic word problems. *Educational Studies in Mathematics*, 103(3), 293–311. <https://doi.org/10.1007/s10649-020-09938-3>
- Shusterman, A., Slusser, E., Halberda, J., & Odic, D. (2016). Acquisition of the cardinal principle coincides with improvement in approximate number system acuity in preschoolers. *PLOS ONE*, 11(4), Article e0153072. <https://doi.org/10.1371/journal.pone.0153072>
- Silliman, D. C., & Kurtz, K. J. (2019). Evidence of analogical re-representation from a change detection task. *Cognition*, 190, 128–136. <https://doi.org/10.1016/j.cognition.2019.04.031>
- Staub, F. C., & Reusser, K. (1995). The role of presentational structures in understanding and solving mathematical word problems. In C. A. Weaver, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension* (pp. 285–305). Lawrence Erlbaum Associates.
- Strohmaier, A. R., MacKay, K. J., Obersteiner, A., & Reiss, K. M. (2020). Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics*, 104(2), 147–200. <https://doi.org/10.1007/s10649-020-09948-1>
- Sulin, R. A., & Dooling, D. J. (1974). Intrusion of a thematic idea in retention of prose. *Journal of Experimental Psychology*, 103(2), 255–262. <https://doi.org/10.1037/h0036846>
- Suppes, P. (1972). *Axiomatic set theory*. Dover Publications.
- Thevenot, C. (2010). Arithmetic word problem solving: Evidence for the construction of a mental model. *Acta Psychologica*, 133(1), 90–95. <https://doi.org/10.1016/j.actpsy.2009.10.004>
- Thevenot, C., & Barrouillet, P. (2015). Arithmetic word problem solving and mental representations. In C. R. Kadosh, & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 158–179). Oxford University Press.
- Trench, M., & Minervino, R. A. (2015). The role of surface similarity in analogical retrieval: Bridging the gap between the naturalistic and the experimental traditions. *Cognitive Science*, 39(6), 1292–1319. <https://doi.org/10.1111/cogs.12201>
- Trench, M., Tavernini, L. M., & Goldstone, R. L. (2017). Promoting spontaneous analogical transfer by idealizing target representations. In R. Granger, U. Hahn, & R. Sutton (Eds.), *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 439–444). Cognitive Science Society.
- Vendetti, M. S., Wu, A., Rowshanshad, E., Knowlton, B. J., & Holyoak, K. J. (2014). When reasoning modifies memory: Schematic assimilation triggered by analogical mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(4), 1172–1180. <https://doi.org/10.1037/a0036350>
- Verschaffel, L. (1994). Using retelling data to study elementary school children's representations and solutions of compare problems. *Journal for Research in Mathematics Education*, 25(2), 141–165. <https://doi.org/10.2307/749506>
- 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. <https://doi.org/10.2307/749836>
- Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM Mathematics Education*, 52(1), 1–16. <https://doi.org/10.1007/s11858-020-01130-4>
- Vicente, S., & Manchado, E. (2016). Arithmetic word problem solving. Are authentic word problems easier to solve than standard ones? Resolución de problemas aritméticos verbales. ¿Se resuelven mejor si se presentan como problemas auténticos? *Infancia y Aprendizaje*, 39(2), 349–379. <https://doi.org/10.1080/02103702.2016.1138717>
- Vicente, S., Verschaffel, L., & Múñez, D. (2021). Comparación del nivel de autenticidad de los problemas aritméticos verbales de los libros de texto españoles y singapurenses [Comparison of the level of authenticity of arithmetic word problems in Spanish and Singaporean textbooks]. *C&E, Cultura y Educación*, 33(1), 106–133. <https://doi.org/10.1080/11356405.2020.1859738>
- 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. <https://doi.org/10.1080/20445911.2014.964719>
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358(6389), 749–750. <https://doi.org/10.1038/358749a0>

(Appendices follow)



## Appendix A

### Materials

**Table A1***Original, Untranslated Problem Statements Used in Experiments 1 and 2*

| Problem identifier | A. Cardinal problems   | Problem identifier | B. Ordinal problems  |
|--------------------|--|--------------------|--|
| Collection 1       | Paul a 5 billes rouges.<br>Il a aussi des billes bleues.<br>En tout, Paul a 14 billes.<br>Charlène a autant de billes bleues que Paul, et des billes vertes.<br>Elle a 3 billes vertes de moins que Paul n'a de billes rouges.<br>En tout, combien Charlène a-t-elle de billes ?                                   | Duration 1         | Le voyage de Sophie dure 8 heures.<br>Son voyage a lieu dans la journée.<br>À son arrivée l'horloge indique 11 heures.<br>Fred part à la même heure que Sophie.<br>Le voyage de Fred dure 2 heures de moins que celui de Sophie.<br>Quelle heure indique l'horloge à l'arrivée de Fred ?   |
| Collection 2       | Sarah a 6 poissons rouges.<br>Ses autres animaux sont des iguanes.<br>Elle possède 15 animaux au total.<br>Bobby garde les iguanes de Sarah pendant les vacances, il les met avec ses tortues.<br>Bobby a 2 tortues de moins que Sarah n'a de poissons rouges.<br>En tout, combien y a-t-il d'animaux chez Bobby ? | Duration 2         | La construction du palais a duré 7 ans.<br>Avant, il a fallu en dessiner les plans.<br>La construction du palais s'est terminée en l'an 15.<br>La construction du château a débuté en même temps que celle du palais.<br>La construction du château a duré 3 ans de moins que celle du palais.<br>En quelle année la construction du château s'est-elle terminée ? |

**Table A2***Ordinal Problem Statements Used in the Solving Task (Original) and in the Recognition Task (Original + Modified) of Experiment 3*

| Problem identifier | Original problem statement  | Modified problem statement  |
|--------------------|---|---|
| Duration 1         | Sofia travelled 5 hours.<br>Her trip started during the day.<br>Sofia arrived at 11 h.<br>Fred left at the same time as Sofia.<br>Fred's trip lasted 2 hours less than Sofia's.<br>What time was it when Fred arrived?  | Sofia travelled 5 hours.<br>Her trip started during the day.<br>Sofia arrived at 11 h.<br>Fred left at the same time as Sofia.<br>Fred arrived 2 hours before Sofia.<br>What time was it when Fred arrived?   |
| Duration 2         | The construction of the palace took 5 years.<br>Plans for the construction were made beforehand.<br>The construction of the palace was completed in year 13.<br>The construction of the castle started at the same time as the construction of the palace.<br>The construction of the castle took 2 years less than the construction of the palace.<br>When was the construction of the castle completed? | The construction of the palace took 5 years.<br>Plans for the construction were made beforehand.<br>The construction of the palace was completed in year 13.<br>The construction of the castle started at the same time as the construction of the palace.<br>The construction of the castle was completed 2 years before the palace.<br>When was the construction of the castle completed? |
| Duration 3         | Rose took painting lessons for 5 years.<br>She started at a certain age.<br>Rose stopped attending painting lessons at age 14.<br>Ted started taking painting lessons at the same age as Rose.<br>He attended painting lessons for 3 years less than Rose.<br>How old was Ted when he stopped attending painting lessons?   | Rose took painting lessons for 5 years.<br>She started at a certain age.<br>Rose stopped attending painting lessons at age 14.<br>Ted started taking painting lessons at the same age as Rose.<br>He stopped attending the lessons 3 years before Rose.<br>How old was Ted when he stopped attending painting lessons?  |
| Elevator 1         | Naomi took the elevator and went up 7 floors.<br>She left from the floor where her grandparents live.<br>She arrived to the 15th floor.<br>Her brother David also took the elevator from their grandparents' floor.<br>He went up 3 floors less than Naomi.<br>What floor did David arrive to?  | Naomi took the elevator and went up 7 floors.<br>She left from the floor where her grandparents live.<br>She arrived to the 15th floor.<br>Her brother David also took the elevator from their grandparents' floor.<br>He arrived 3 floors below Naomi.<br>What floor did David arrive to?  |

*(table continues)**(Appendices continue)*

**Table A2** (*continued*)

| Problem identifier | Original problem statement  | Modified problem statement  |
|--------------------|---|---|
| Elevator 2         | Katherine took the elevator and went up 7 floors.<br>She left from the floor where the gym is.<br>She arrived to the 13th floor.<br>Yoan also took the elevator from the floor where the gym is.<br>He went up 2 floors less than Katherine.<br>What floor did Yoan arrive to?  | Katherine took the elevator and went up 7 floors.<br>She left from the floor where the gym is.<br>She arrived to the 13th floor.<br>Yoan also took the elevator from the floor where the gym is.<br>He arrived 2 floors below Katherine.<br>What floor did Yoan arrive to?  |
| Elevator 3         | Gloria took the elevator and went up 5 floors.<br>She left from the floor where her office is.<br>She arrived to the 14th floor.<br>Her coworker Larry also took the elevator from their office's floor.<br>He went up 2 floors less than Gloria.<br>What floor did Larry arrive to?  | Gloria took the elevator and went up 5 floors.<br>She left from the floor where her office is.<br>She arrived to the 14th floor.<br>Her coworker Larry also took the elevator from their office's floor.<br>He arrived 2 floors below Gloria.<br>What floor did Larry arrive to?  |
| Height 1           | Slouchy Smurf is 6 centimeters tall.<br>He climbs on a smurf table.<br>He now attains the height of 11 centimeters.<br>Grouchy Smurf climbs on the same table as Slouchy Smurf.<br>Grouchy Smurf is 2 centimeters shorter than Slouchy Smurf.<br>What height does Grouchy Smurf attain when he climbs on the table?                     | Slouchy Smurf is 6 centimeters tall.<br>He climbs on a smurf table.<br>He now attains the height of 11 centimeters.<br>Grouchy Smurf climbs on the same table as Slouchy Smurf.<br>The height Grouchy Smurf attains on the table is 2 centimeters shorter than that of Slouchy Smurf.<br>What height does Grouchy Smurf attain when he climbs on the table?                 |
| Height 2           | Obelix's statue is 5 meters tall.<br>It is placed on a pedestal.<br>Once on the pedestal, it attains 12 meters in height.<br>Asterix's statue is placed on the same pedestal as Obelix's.<br>Asterix's statue is 3 meters shorter than Obelix's.<br>What height does Asterix's statue attain when placed on the pedestal?               | Obelix's statue is 5 meters tall.<br>It is placed on a pedestal.<br>Once on the pedestal, it attains 12 meters in height.<br>Asterix's statue is placed on the same pedestal as Obelix's.<br>The height Asterix's statue attains on the pedestal is 3 meters shorter than that of Obelix's statue.<br>What height does Asterix's statue attain when placed on the pedestal? |
| Height 3           | The giraffe in the zoo is 6 meters tall.<br>It climbs on the biggest rock in the park.<br>Once on the rock, it attains 11 meters in height.<br>The elephant in the zoo climbs on the same rock as the giraffe.<br>The elephant is 3 meters shorter than the giraffe.<br>What height does the elephant attain when standing on the rock? | The giraffe in the zoo is 6 meters tall.<br>It climbs on the biggest rock in the park.<br>Once on the rock, it attains 11 meters in height.<br>The elephant in the zoo climbs on the same rock as the giraffe.<br>The height he attains on the rock is 3 meters shorter than the giraffe's.<br>What height does the elephant attain when standing on the rock?              |

*Note.* Changes were systematically introduced in the fifth sentence. The problems' numerical values varied between participants.

(*Appendices continue*)

**Table A3***Cardinal Problem Statements Used in the Solving Task (Original) and in the Recognition Task (Original + Modified) of Experiment 3*

| Problem identifier | Original problem statement  | Modified problem statement   |
|--------------------|---|--|
| Collection 1       | Paul has 7 red marbles.<br>He also has blue marbles.<br>In total, Paul has 13 marbles.<br>Jolene has as many blue marbles as Paul, and some green marbles.<br>She has 2 green marbles less than Paul has red marbles.<br>How many marbles does Jolene have?   | Paul has 7 red marbles.<br>He also has blue marbles.<br>In total, Paul has 13 marbles.<br>Jolene has as many blue marbles as Paul, and some green marbles.<br>She has 2 marbles less than Paul.<br>How many marbles does Jolene have?  |
| Collection 2       | Sarah owns 7 goldfish.<br>Her other pets are all iguanas.<br>In total, she owns 15 pets.<br>Bobby is pet-sitting Sarah's iguanas during the holidays, he puts them with his pet turtles.<br>Bobby owns 3 turtles less than Sarah owns goldfish.<br>How many pets are there at Bobby's?  | Sarah owns 7 goldfish.<br>Her other pets are all iguanas.<br>In total, she owns 15 pets.<br>Bobby is pet-sitting Sarah's iguanas during the holidays, he puts them with his pet turtles.<br>Now at Bobby's, there are 3 pets less than there were at Sarah's before the holidays.<br>How many pets are there at Bobby's?   |
| Collection 3       | Karl picked 6 tulips.<br>He puts them with the daffodils he gathered.<br>In total, Karl has 11 flowers in his bouquet.<br>In her bouquet, Clarice has as many daffodils as Karl, and some roses.<br>She has 2 roses less than Karl has tulips.<br>How many flowers does Clarice have in her bouquet?  | Karl picked 6 tulips.<br>He puts them with the daffodils he gathered.<br>In total, Karl has 11 flowers in his bouquet.<br>In her bouquet, Clarice has as many daffodils as Karl, and some roses.<br>Clarice has 2 flowers less than Karl does.<br>How many flowers does Clarice have in her bouquet?   |
| Price 1            | In the store, Anthony wants to buy a 5-dollar ruler.<br>He also wants a notebook.<br>In total, that will cost him 13 dollars.<br>Julie wants to buy the same notebook as Anthony, and an eraser.<br>The eraser costs 2 dollars less than the ruler.<br>How much will Julie have to pay?   | In the store, Anthony wants to buy a 5-dollar ruler.<br>He also wants a notebook.<br>In total, that will cost him 13 dollars.<br>Julie wants to buy the same notebook as Anthony, and an eraser.<br>In total, she will pay 2 dollars less than Anthony will.<br>How much will Julie have to pay?   |
| Price 2            | The first meal on the menu includes a chocolate cake costing 5 dollars.<br>The meal also includes a mushroom omelet.<br>In total, that makes for a 14-dollar meal.<br>The second meal on the menu includes the same mushroom omelet, and an apple pie.<br>The apple pie costs 3 dollars less than the chocolate cake.<br>How much does the second meal cost?                                    | The first meal on the menu includes a chocolate cake costing 5 dollars.<br>The meal also includes a mushroom omelet.<br>In total, that makes for a 14-dollar meal.<br>The second meal on the menu includes the same mushroom omelet, and an apple pie.<br>The second meal on the menu is 3 dollars cheaper than the first meal.<br>How much does the second meal cost?                   |
| Price 3            | Tyler wants to buy French fries that cost 5 dollars.<br>He will also take a cheeseburger.<br>In total, that will cost him 14 dollars.<br>Zoey orders a cheeseburger as well, and a milkshake.<br>The milkshake costs 2 dollars less than the French fries.<br>How much will Zoey pay for her order?   | Tyler wants to buy French fries that cost 5 dollars.<br>He will also take a cheeseburger.<br>In total, that will cost him 14 dollars.<br>Zoey orders a cheeseburger as well, and a milkshake.<br>Her order will cost 2 dollars less than Tyler's.<br>How much will Zoey pay for her order?   |
| Weight 1           | Joe takes a Russian dictionary weighing 5 kilograms.<br>He also takes a Spanish dictionary.<br>In total, he is carrying 12 kilograms of books.<br>Lola takes Joe's Spanish dictionary and a German dictionary.<br>The German dictionary weighs 3 kilograms less than the Russian dictionary.<br>How many kilograms of books is Lola carrying?   | Joe takes a Russian dictionary weighing 5 kilograms.<br>He also takes a Spanish dictionary.<br>In total, he is carrying 12 kilograms of books.<br>Lola takes Joe's Spanish dictionary and a German dictionary.<br>In total, Lola's books weigh 3 kilograms less than Joe's.<br>How many kilograms of books is Lola carrying?   |
| Weight 2           | A bag of pears weighs 6 kilograms.<br>It is weighed together with cheese.<br>In total, the weighing scale indicates 11 kilograms.<br>The same cheese is then weighed together with a milk carton.<br>The milk carton weighs 3 kilograms less than the bag of pears.<br>What is indicated on the weighing scale now?   | A bag of pears weighs 6 kilograms.<br>It is weighed together with cheese.<br>In total, the weighing scale indicates 11 kilograms.<br>The same cheese is then weighed together with a milk carton.<br>In total, the weighing scale indicates 3 kilograms less than before.<br>What is indicated on the weighing scale now?  |
| Weight 3           | On moving day, Ryan is carrying his microwave oven, which weighs 5 kilograms.<br>He is carrying his coffee machine at the same time.<br>In total, he is carrying 11 kilograms of appliances.<br>Felicia takes Ryan's coffee machine from him while carrying a blender.<br>The blender weighs 2 kilograms less than the microwave oven.<br>How many kilograms of appliances is Felicia carrying? | On moving day, Ryan is carrying his microwave oven, which weighs 5 kilograms.<br>He is carrying his coffee machine at the same time.<br>In total, he is carrying 11 kilograms of appliances.<br>Felicia takes Ryan's coffee machine from him while carrying a blender.<br>In total, she is carrying 2 kilograms less than Ryan.<br>How many kilograms of appliances is Felicia carrying? |

*Note.* Changes were systematically introduced in the fifth sentence. The problems' numerical values varied between participants.

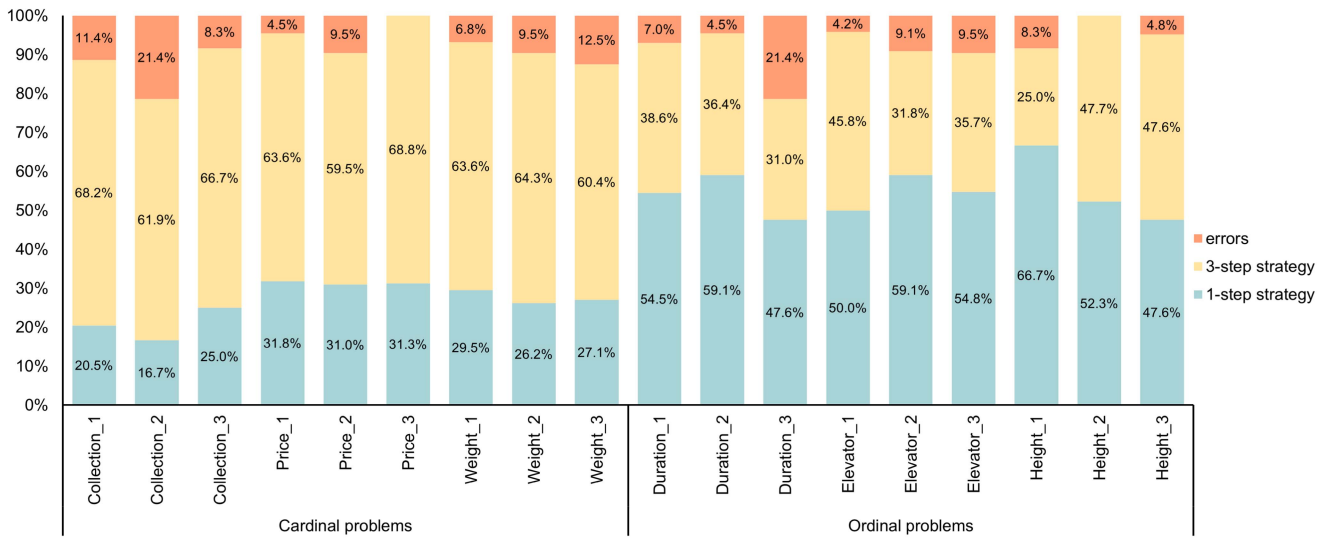
(Appendices continue)

## Appendix B

### Supplementary Analysis of Experiment 3

**Figure B1**

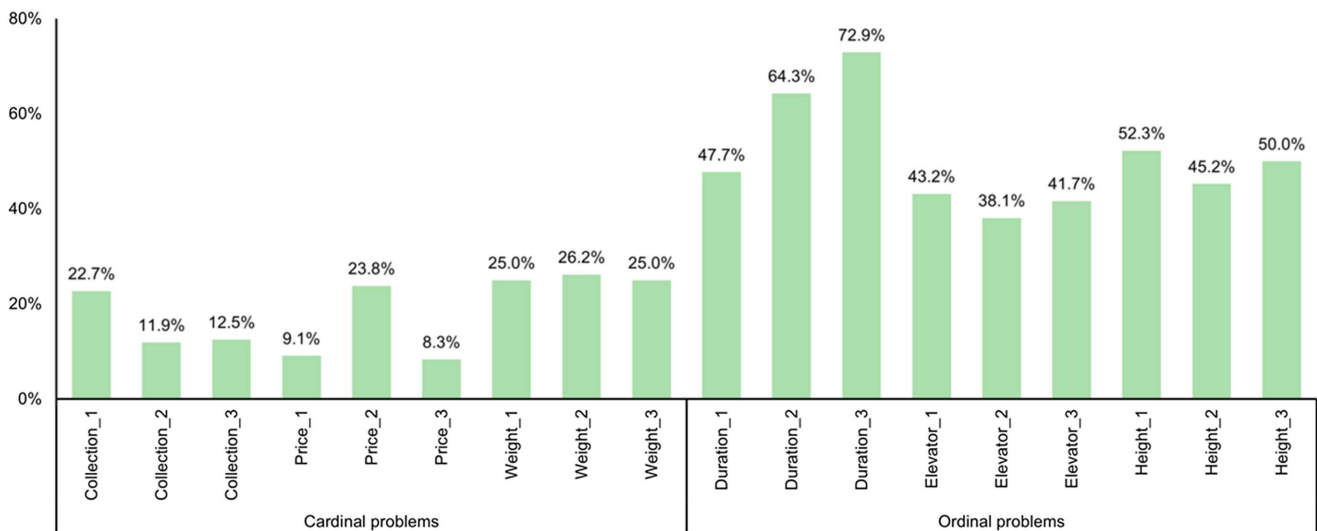
*Percentage of Use of the One-Step Strategy, the Three-Step Strategy, as Well as Error Rates for Each Problem Statement in Experiment 3*



*Note.* See the online article for the color version of this figure.

**Figure B2**

*Percentage of False Recognition of Altered Sentences Revealing a Misremembrance of the Part-to-Part Difference as a Whole-to-Whole Difference Instead, for Each Problem Statement in Experiment 3*



*Note.* See the online article for the color version of this figure.

*(Appendices continue)*



Appendix C  
Control Experiment

To investigate whether participants' strategy choice on cardinal and ordinal problems could be attributed to variations in working memory demand, we designed a control experiment contrasting two different wordings of cardinal problems. Specifically, we aimed to determine whether the mention of entities differing by their physical properties (e.g., counting *green* marbles vs. *red* marbles in Problem C1) could account for some of the difficulty experienced by participants when attempting to find the one-step strategy on cardinal problems. Indeed, ordinal problems tend to mention entities associated with different protagonists (e.g., floors traveled by *Paul* vs. floors traveled by *Gloria* in Problem E3) instead of entities possessing different physical properties, which may induce a lower load in working memory. To assess whether this difference constitutes a potential confound and to gain a more detailed understanding of how the properties of entities influence participants' problem representations, we compared the performance of 100 participants on two target problems.

Method

Participants

We recruited 100 participants ( $M_{\text{age}} = 43.27$  years,  $SD = 9.08$ ) on Amazon Mechanical Turk to complete this online experiment. Participants were required to be fluent English speakers. They were paid \$1.00 USD for their participation and had to complete the task within 1 hr. Participants' compensation was funded by a grant from the CY Initiative of Excellence (CYIn-AAP2021-AmbEm-0000000026).

Materials and Procedure

The experiment was hosted on the Qualtrics platform for online experiments. Each participant was instructed to solve two problems, using as few operations as possible (the instructions were similar to the ones in Experiment 3). The goal of the experiment was to compare participants' performance on two different versions of the cardinal problem: one was the Collection 1 problem used in Experiment 3 (the marble problem involving a comparison between

red and green marbles). The other one was a modified version of this problem, in which we had replaced the fourth and fifth sentences to mention red marbles instead of green marbles, so that Paul's and Jolene's extra marbles would not differ based on their color. Table C1 presents the two versions side by side:

In the new version of the Collection 1 problem, Paul and Jolene's extra marbles have the same color, which means that the difference between Part 1 and Part 3 is not a difference of physical properties but a difference of whom they belong to (similar to how Gloria's and Larry's elevator trips differ in who was traveling, in Problem E3). The goal was to compare the two problems and identify whether the new version would lead to a higher percentage of use of the one-step strategy, closer to the percentage observed on ordinal problems in Experiments 1 to 3.

Each participant was randomly assigned one of the two collection problems by the Qualtrics platform. They had to solve the cardinal problem first and were then all presented with the same ordinal problem (Elevator 3) to solve, for comparison.

Results

We coded participants' responses depending on whether they solved the problems using the one-step strategy, the three-step strategy, or whether they made an error and failed to solve the problems. The descriptive results are presented in Figure C1.

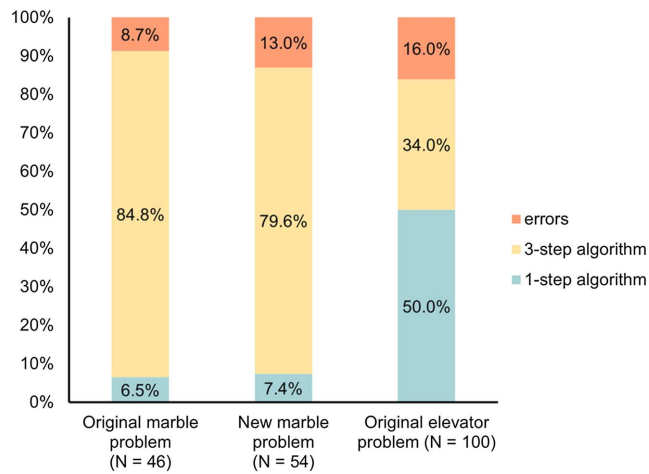
Interestingly, participants' propensity to find and use the one-step algorithm did not noticeably change between the original version of the marble problem (6.5%) and the new version of the marble problem (7.4%). On the other hand, participants' propensity to find the one-step algorithm on the elevator problem remained considerably higher (50.0%) than both marbles problems. Thus, modifying the marble problem so that the difference between Part 1 and Part 3 would be presented as a difference between Paul's red marbles and Jolene's red marbles (instead of between Paul's red marbles and Jolene's green marbles) did not appear to substantially change participants' solving strategies. This result suggests that the mention of entities having different properties in some of the cardinal problems could not alone account for the representational differences between cardinal and ordinal problems in Experiments 1 to 3.

Table C1  
The Two Versions of the Marble Problem (Collection 1) Used in the Control Experiment

| Collection Problem 1 (original)   | Collection Problem 1 (new version)  |
|---|---|
| Paul has 7 red marbles.<br>He also has blue marbles.<br>In total, Paul has 13 marbles.<br>Jolene has as many blue marbles as Paul, and some green marbles.<br>She has 2 green marbles less than Paul has red marbles.<br>How many marbles does Jolene have? | Paul has 7 red marbles.<br>He also has blue marbles.<br>In total, Paul has 13 marbles.<br>Jolene has as many blue marbles as Paul, and some red marbles.<br>She has 2 red marbles less than Paul.<br>How many marbles does Jolene have? |

Note. The problems differed in their fourth and fifth sentence. (Changes highlighted in red for convenience—those highlights were not present in the version presented to the participants.)

(Appendices continue)

**Figure C1***Distribution of Solving Strategies and Errors in the Control Experiment*

*Note.* Participants were randomly assigned either the original version of the marble problem ( $N = 46$ ) or the new version ( $N = 54$ ). They all had to solve the elevator problem afterward ( $N = 100$ ). See the online article for the color version of this figure.

Received May 3, 2023

Revision received April 11, 2024

Accepted April 21, 2024 ■