Please cite the published version:

Sroka, M., Obidziński, M., Cipora, K., & Hohol, M. (2026). Spatial Biases in Arithmetic: The Effect of Operation Order and Visual Spacing on Calculations. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218261490417

Spatial biases in arithmetic: The effect of operation order and visual spacing on calculations

Martyna Sroka*, Michał Obidziński, Krzysztof Cipora$, Mateusz Hohol$

Mathematical Cognition and Learning Lab, Copernicus Center for Interdisciplinary Studies, Jagiellonian University, Krakow, Poland

* Corresponding author: Martyna Sroka, Mathematical Cognition and Learning Lab, Copernicus Center for Interdisciplinary Studies, Jagiellonian University, Main Square 34 (Spiski Palace, fl 3, rm 1), 31-010, Krakow, Poland, e-mail: sroka.martyna@outlook.com
$ Equal contributions

ORCIDs:
MS: 0009-0006-2387-4175
MO: 0000-0002-7854-3123
KC: 0000-0003-0077-9336
MH: 0000-0003-0422-5488

Statements and Declarations

Ethical considerations

The study was approved by the Rector's Committee for Research Ethics of the Jagiellonian University (1027.0041.1.2025). All participants provided their informed consent in an online form.

Declaration of conflicting interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding statement

This scientific study was funded by the state budget through the Ministry of Science and Higher Education of Poland under the program "Science for Society II", project ID: NdS-II/SN/0332/2024/01; the total value of the project 1 000 000 PLN.

Data availability

The dataset and analysis code are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/EZYSB).

Abstract

Visual layout is known to affect computational performance, yet its relation to response format and numerical proficiency remains unclear. This preregistered study examined whether spatial presentation modulates arithmetic problem-solving. We conceptually replicated Experiment 3 of Landy and Goldstone's (2010) study using the Precedence and Proximity Task in a forced-choice format. Additionally, we assessed arithmetic fluency using a speeded test. A total of 201 adults solved 200 mixed addition–multiplication expressions in which spacing either aligned or conflicted with order-of-operations rules. Operation order was also manipulated. Participants were faster and more accurate in congruent spacing conditions, whereas incongruent spacing increased operation-order errors, especially when problems began with addition. A subgroup of participants showed disproportionately poor performance under incongruent spacing, suggesting a prepotent susceptibility to misleading grouping cues. The spacing effect persisted despite differences in task format and response procedure compared to the original study, supporting its robustness. Higher arithmetic fluency was associated with higher accuracy and shorter reaction times in the Precedence and Proximity Task. These findings corroborate that visuospatial layout biases symbolic computation and that individuals differ in their susceptibility to it.

Keywords: perceptual grouping; spacing effect; mathematical cognition; number processing

Introduction

Perceptual Influences in Rule-Based Tasks

Humans often navigate the world using both formal rules (explicitly taught, codified, involving abstract reasoning) and informal rules (learned through shared practice and convention). Importantly, reasoning does not rely solely on formally acquired knowledge. The human perceptual system continuously extracts structure from the environment, and perceptual cues can guide interpretation and decision-making — even when they are formally irrelevant (Henik & Tzelgov, 1982). In some cases, such cues can bias or even override explicitly taught rules, revealing an interaction between abstract symbolic knowledge and perceptual processing. Studying when and how such perceptual influences interfere with rule-governed behavior provides insight into the mechanisms underlying human information processing.

Numerical cognition is an excellent field to look into the processing of abstract conventions (Campbell, 2015; see also Poincaré, 1952). To deal with mathematics, one needs to learn and follow a specific set of rules, among which the order of arithmetic operations is fundamental. However, applying abstract rules is difficult for many individuals and thus prone to errors. Adherence to the order of operations is not free from the influence of more general ways in which people process information (Landy et al., 2014; Landy & Goldstone, 2010).

Spatial–Numerical Associations and Perceptual Grouping

Spatial processing appears to be tightly linked with numerical processing (Bueti & Walsh, 2009; Cipora et al., 2018; Cipora, Haman, et al., 2020; Hawes & Ansari, 2020). This link is often referred to as Spatial-Numerical Associations (SNAs). Among many categories of such associations, extension SNAs can be distinguished, reflecting how numerical information is mapped onto spatial representations (Cipora et al., 2015). Within this category, the approximate mapping appears in tasks involving physical and numerical size and spacing manipulation (Henik & Tzelgov, 1982; Lonnemann et al., 2008; Vogel et al., 2019).

Spacing and Operator Precedence

General principles of perception, as described by Gestalt psychologists, can also interfere with the processing of abstract information (Hohol et al., 2017). In particular, perceptual grouping — especially the law of proximity (Wertheimer, 1922/2017) — may affect the interpretation of abstract problems, depending on their visual layout. The perceptual properties of notations and visual details in the presentation of formal languages may influence how people reason and analyze data (Kirshner, 1989; Landy & Goldstone, 2007a). A notable demonstration of this interaction was provided by Landy and Goldstone (2007b), who showed that the accuracy of algebraic problem-solving is affected by perceptual grouping biases, such as spacing, similarity, connectedness, and common region. In their later work (Landy & Goldstone, 2010), the authors demonstrated across a series of three experiments using open-format responses that manipulating the spacing between operands in addition and multiplication expressions systematically biases how arithmetic expressions are parsed and solved. In Experiment 1, participants solved single-operation problems (addition or multiplication) presented with either narrow or wide spacing, revealing that overall spatial extent influenced numerical responses and operation confusions. Experiment 2 extended this to compound expressions (plus–plus, plus–times, times–plus, times–times) with varying spacing patterns, showing that participants were more likely to commit order-of-operations errors when spacing was incongruent with operator precedence. Experiment 3 focused exclusively on mixed-operation expressions (plus–times and times–plus; 200 expressions) and demonstrated that spacing congruency reliably affected both accuracy and response times, even though spacing is formally irrelevant to the correct solution. This effect emerged even though participants were reminded of the rules governing arithmetic operations. Importantly, spacing manipulation was relatively large and arguably less naturalistic than the spacing typically encountered in everyday arithmetic notation. Overall, participants tended to solve the visually grouped parts of the problems first, even when doing so violates the standard order of operations.

Replication and Study Rationale

This preregistered online study aimed to conceptually replicate the third experiment reported in the abovementioned article *Proximity and precedence in arithmetic* (Landy & Goldstone, 2010). Specifically, we investigated how physical spacing between operands affects the calculation of mixed addition and multiplication expressions. In contrast to the production-format paradigm used in the original study, we employed a forced-choice response format with four answer options. From now on, we refer to this task as the "Precedence and Proximity Task". We aimed to check the robustness of the effect when the stimuli are presented in multiple choice format (to our best knowledge never done before). Prior studies have predominantly used production paradigms in which participants generate and type complete solutions (Braithwaite et al., 2016; Harrison et al., 2020; Jiang et al., 2014; Rivera & Garrigan, 2016)

Beyond replication, we were interested in a possible association between perceptual grouping biases and arithmetic fluency in adults, measured with the Math4Speed task (M4S, Loenneker et al., 2025). Developmental research has shown that this kind of perceptual bias has been stronger in older pupils (equivalent of 5th-12th grade in USA) than in young ones (2nd-6th grade) (Braithwaite et al., 2016). Therefore, it is possible that the level of arithmetic skills in adults is also related to the occurrence and strength of this effect. It is worth investigating whether this bias varies in adults with different levels of arithmetic competence. On the one hand, the level of mathematical skills should make individuals more reliant on their knowledge of formal rules and less susceptible to distractions from irrelevant information (Ding et al., 2024; Kelley & Yantis, 2009). Therefore, we would expect participants who are more arithmetically fluent to be less susceptible to these spatial biases and, consequently, to perform better (in terms of accuracy and reaction times, RTs) on the Precedence and Proximity Task. On the other hand, in line with research by Braithwaite and colleagues (2016), the grouping bias may be stronger due to previously learned notation conventions — such as interpreting grouped operands as if they were enclosed in brackets.

Previous Research

The findings of Landy and Goldstone (Goldstone et al., 2010; Landy & Goldstone, 2010) have been widely cited as evidence that symbolic notation and Gestalt principles interfere with arithmetic problem-solving (Abrahamson et al., 2020). Their results serve, among other research, as a foundation for math training programs (Decker-Woodrow et al., 2023), perceptual learning-based interventions (Ottmar et al., 2015), and even proposed changes in educational policies and systems (Goldstone et al., 2010). Moreover, they carry important theoretical implications for spatial-numerical associations (SNAs), as they represent a core phenomenon within one of the SNA categories (Cipora et al., 2018). Given this broad impact, we aimed to test whether Landy and Goldstone's results hold under modified response format and more natural spatial arrangement. Only a handful of similar experiments have examined this kind of spatial bias, including manipulations of spacing and the position of minus signs in expressions, the use of mental arithmetic tasks with visual masking and encouraged verbalization (Rivera & Garrigan, 2016), and studies with children designed to investigate developmental and learning-related effects in real-world online learning environments (Braithwaite et al., 2016; Harrison et al., 2020). Jiang et al.(2014) demonstrated that altering the position of minus signs and manipulating spacing increased error rates, suggesting that participants relied on perceptual rather than formal mathematical groupings. Developmental research by Braithwaite et al. (2016) found that younger children's (8-12 years old, equivalent of 2nd-6th graders in USA) susceptibility to narrow spacing increased with grade level, while Harrison et al. (2020) showed that older students (5th-12th graders) were more error-prone and performed worse when solving problems with incongruent spacing, regardless of age or prior ability. Altogether previous research showed a tendency for increased errors in incongruent conditions (narrowly spaced addition and widely spaced multiplication). Importantly, all of these studies used a production response format, which limits the precision of RT measurement. A detailed overview of prior studies can be found in Supplementary Table S1. Here, we focus specifically on the relationship between precedence rule violations and the spacing between operands, which Landy and Goldstone (2010) investigated in their Experiment 3, and we modified the original procedure in a number of ways.

Research questions

Based on this rationale, the current study addressed the following questions. Firstly, we tested whether the effect holds in a forced-choice task. Landy and Goldstone (2010), as well as authors of the replications mentioned above, used a production paradigm, measuring RTs from the first key press for correct answers. Replacing the production paradigm with a forced-choice response format allowed us to investigate the robustness of the effect of interest on a shorter timescale and with more precise RTs' measurements than in the original study. Each time, we presented four response alternatives and required participants to press a corresponding response key. This methodological choice allowed us to capture the moment when participants were ready to make their final answer. Our approach contrasts with measuring the moment when typing begins — when the answer may still change during the process, especially when the response is a multi-digit number. We chose a four-choice response format rather than a correct/incorrect decision task to minimize reliance on estimation strategies and to require exact calculation. Importantly, our selection of possible answers always included an answer for a specific type of error — an operation-order error — as an alternative to the correct answer and two distractor answers. One of the main conclusions from prior studies was susceptibility to operation-order errors in incongruently spaced problems. Furthermore, this modification aimed to facilitate the procedure in an online testing environment, given the variability in keyboards and typing skills. Finally, the multiple-choice format is widely used to present mathematical problems in school and university exams and was extensively adopted in online testing environments during the pandemic. Therefore, our approach reflects an important aspect of the current educational landscape.

Secondly, we adopted a more naturalistic spacing manipulation than that used by Landy and Goldstone (2010) (i.e., 18.2 cm total width; 4 cm between narrowly spaced operands, 10 cm between widely spaced ones, with symbols 1.4 cm wide). While preserving the original narrow-to-wide spacing ratio, we reduced the absolute distance between operands, using 64 px for narrowly spaced problems and 160 px for widely spaced problems (see Fig S1 in Supplementary Materials). This modification aimed to determine whether spacing effects on problem-solving occur robustly in more naturalistic settings, where the presentation format more closely resembles the typical formatting of arithmetic problems in mathematics classes. Based on the abovementioned replication studies (c.f. Table S1 in Supplementary Materials), which used more natural spacing, we assumed the effect should still be present in such layouts.

Thirdly, as we used the original stimulus set design, we omitted the problem size analysis. Even though Landy and Goldstone (2010) considered the problem size effect, the stimulus set they used was unbalanced in terms of problem size. They defined problem size by the magnitude of the largest operand. In the original study, there were five possible values for the outer operands (2, 3, 6, 8, 9), which were classified as either small (2, 3) or large (6, 8, 9). However, the middle operand — always 3 or 4 — was not analyzed, despite indicating a small problem size. In the original Experiment 3, participants had higher operand and operation error proportion for larger operands, but specifically precedence errors were affected by spacing rather than by operand magnitude (Landy & Goldstone, 2010). In our study, the experimental design was balanced with respect to congruency (congruent vs. incongruent spacing) and the order of operations in the equations (addition-first vs. multiplication-first), but the issue of problem size remained unchanged. Authors of other replication studies chose a similar approach.

Fourthly, due to the nature of online data collection, we were unable to control for viewing distance (similarly to majority of the replications). In the original experiment, the viewing distance was set at 55 cm, in Rivera and Garrigan (2016) experiments it was set at 122 cm. Instead, we chose to display the stimuli in px and allow relative scaling.

Finally, we included a speeded test of arithmetic fluency and asked participants about their educational level and school math performance, while previous follow-ups of Landy and Goldstone's (2010) study did not directly measure participants' math skills. Moreover, in addition to traditional group-level analyses, we introduced a novel analytical component by adopting an individual-level approach, following Grice et al. (2020), who proposed treating "persons as effect size." This method complements group-level statistics by quantifying the percentage of participants whose performance pattern aligns with the theoretical prediction. Reporting this percentage provides an indication of how consistently the observed group effect is reflected across individuals.

In line with Landy and Goldstone's (2010) findings, narrow spacing between operands in multiplication and wide spacing in addition (congruent condition, e.g., [2 × 2   +   3] and [3   +   2 × 2]) is expected to facilitate the task performance in terms of better accuracy and shorter RTs, as the perceptual grouping aligns with the order of operations. In contrast, changing the spacing to present operands closer in addition and more distant in multiplication (incongruent condition, e.g., [2   ×   2 + 3] or [3 + 2   ×   2]) should make participants more susceptible to making operation-order errors (e.g., responding 10 to the problem 3 + 2 × 2). In addition to the effects of spacing congruency on arithmetic performance, the order in which operations are presented in the problem is likely to influence participants' error rates and response speed (Landy & Goldstone, 2010). Problems that begin with addition rather than multiplication are expected to generate greater cognitive load, as they require participants from left-to-right reading cultures to override their natural tendency to process visual material and count from left to right (Shaki et al., 2012), potentially leading to longer RTs and more operation-order errors.

Preregistered hypotheses

H1 Congruency

Congruent spacing (larger gaps for addition and narrower ones for multiplication) is expected to facilitate calculation, resulting in higher accuracy and shorter RTs compared to incongruent spacing, in which visual grouping cues conflict with the order of operations. Moreover, we expect the proportion of operation-order errors to be higher in the incongruent condition

H2 Operation Order

Solving problems that begin with addition takes longer and is more likely to result in operation-order errors, leading to lower overall accuracy. We expect the proportion of operation-order errors to be higher in the case of problems that start with addition

H3 Arithmetic fluency and precedence in arithmetic

Higher fluency in basic arithmetic, measured by the M4S, is expected to correlate with better performance on the Precedence and Proximity Task, resulting in greater accuracy (preregistered) and, additionally to what we preregistered, shorter RTs. We expected to find interaction effects between Congruency and Operation Order, however, we did not formulate explicit hypotheses regarding their shape or direction.

Materials and Method

Participants

The final analytic sample consisted of 201 individuals (mean age *M* = 25.4, *SD* = 3.3 years; 99 women, 100 men, 2 participants preferred not to disclose their gender). The group included mostly university students or graduates: students without a degree (*n* = 57), bachelor's degree holders (*n* = 75), individuals with completed or ongoing master's-level education (*n* = 31), PhD students or graduates (*n* = 6). Other participants were high school graduates (n = 31), and one individual had less than a high school education. Participants represented different fields of studies: STEM (*n* = 47), Arts and Humanities (*n* = 39), Medicine/Medical Sciences (*n* = 30), other/not applicable (*n* = 85).

Participants were recruited online via Prolific and received £7 for participation. A total of 221 complete responses were collected. Because three participants completed the task twice, only their first attempt was retained, yielding an initial sample of 218 English-fluent adults aged 18–30 years. Of these, 17 participants were excluded from the final analysis — due to performance below chance level (25%) and/or consistently short RTs (less than 500 ms for at least 15% of responses), suggesting low-effort responding (e.g., random answer selection without attempting to solve the task).

The preregistered target sample size was set at a minimum of 200 participants, based on an a priori power analysis. This ensured sufficient statistical power (>.80, with α = .05) to detect a within-subject effect size (Cohen's *d*) of 0.2 and a Pearson correlation of .2, which were defined as the smallest effects of interest.

A small group of participants (*n* = 25) in the final analytic sample exhibited atypical performance patterns, with near-zero accuracy in at least one condition despite overall accuracy above the chance level (> 25%). Most of these individuals (21 out of 25) demonstrated an understanding of the precedence rule, as reflected in above-chance accuracy in the addition-first condition and were therefore retained in the analysis. In contrast, four participants scored below chance across all addition-first trials, suggesting a possible misunderstanding of operation order rule. These cases were identified but not excluded considering that the precedence rule was explicitly reminded at the beginning of the task. However, to ensure that the overall results are not driven by these individuals, we additionally ran the main analysis again excluding them.

Procedures and Materials

Participants completed the Precedence and Proximity Task, a speeded test of arithmetic fluency, and a demographic survey including questions about their math grades. The task order was fixed. The median duration time of the entire procedure was 36 minutes. At the beginning of the session, participants were presented with the following instructions:

> "*This study examines how the way math problems are presented affects how well people solve them. The study consists of three parts.*
> *-In the first part, your task will be to choose the correct answer in a single-choice test.*
> *-In the second part, your task will be to provide the result of an arithmetic calculation.*
> *-Finally, we will ask you to complete a short demographic questionnaire.*
> *Complete the tasks in one go without breaks. We will let you know when it's time to rest. The estimated participation time is about 45 minutes.*"

Precedence and Proximity Task

Each participant was required to solve a centrally presented arithmetic equation and respond in a forced-choice format with four possible answers by pressing one of four keys (D, F, J, or K) as quickly and accurately as possible. No time limit was imposed. The arrangement of response options on the screen corresponded to the order of the keys on the keyboard (e.g., to select the leftmost answer, participants pressed D; to select the third one, J). To familiarize participants with the method of choosing answers using the keyboard, each session began with a practice session consisting of eight equations (participants received the following instruction:

"In this experiment, your task is to solve the arithmetic problem and choose the correct answer from four options. Place your fingers on the keyboard as follows: – Left middle finger on the D key, – Left index finger on the F key, – Right index finger on the J key, – Right middle finger on the K key. In each trial, a fixation dot will appear at the center of the screen. Four possible answers will then be displayed. Press the key that corresponds to the correct answer. Please respond as quickly and accurately as possible. The experiment will begin with a practice session, during which you will receive feedback on your accuracy. Remember to complete the task in one sitting without taking breaks.").

The practice session contained only addition or multiplication problems with three operands (no mixed-operation problems). Participants received accuracy feedback after each response. The test session began with a reminder of the rules of operation order:

"You are about to begin the main task. It will involve solving arithmetic problems using addition and multiplication. Remember to follow the rule for the order of operations: always perform multiplication before addition. For example, the correct answer to 2 + 2 × 2 is 6."

It was followed by the instruction:

"The first part of the experiment will begin now. Please keep your fingers on the same keys (D, F, J, and K) and give answers to the equations analogically to the practice session. Please answer as quickly and as accurately as possible. The feedback on the correctness of your answers will not be displayed from this point onward. Press SPACE on your keyboard to start."

In test session participants solved 200 arithmetic expressions involving mixed operations – additions and multiplications, in which the spacing between operands was manipulated. Each participant completed a single session at a time and place of their choosing, with the only requirement that the procedure was conducted on a personal computer (desktop or laptop, i.e., a device with a physical QWERTY keyboard was required).

Table 1

Stimulus Types and Example Trial Display

a. Types of stimuli

	Congruent Spacing	**Incongruent Spacing**
Addition first	**a + b × c** e.g. **6 + 4 × 8**	**a + b × c** e.g. **6 + 4 × 8**
Multiplication first e.g.	**a × b + c** e.g. **6 × 4 + 8**	**a × b + c** e.g. **6 × 4 + 8**

Note. Stimuli crossed Operation Order (addition-first vs. multiplication-first) with Spacing Congruency (congruent vs. incongruent). In the addition-first format, expressions were of the form a + b × c; in the *multiplication-first format, expressions were of the form a × b + c. Congruent spacing displayed the multiplication operands closer together than the addition operands, whereas incongruent spacing displayed the addition operands closer together than the multiplication operands. The stimulus set consisted of 200 trials (4 conditions × 50 trials). The second operand (b) was always 3 or 4. The first and third operands (a and c) were sampled from 2, 3, 6, 8, and 9.*

b. Example of a stimulus with incongruent spacing: arithmetic expression with four answers in forced-choice format

Arithmetic expression	**6 + 4 × 8**			
Response alternative	**40**	**78**	**38**	**80**
Response key	**D**	**F**	**J**	**K**

Note. In the example trial, the correct answer was 38 (J). The "operation-order error" alternative (i.e., computing addition before multiplication) was 80 (K). The remaining two alternatives (40 [D] and 78 [F]) were distractors selected to be within ±2 of either the correct answer or the operation-order error answer; depending on the trial, distractors could both be larger, both be smaller, or be mixed relative to these anchors.

The test stimulus set consisted of 200 arithmetic expressions (e.g.: [a × b + c] or [a + b × c]; see Table 1a). The design manipulated two factors: Operation Order (plus–times vs. times–plus) and Spacing Congruency. In the plus–times format, addition preceded multiplication (e.g. [a + b × c]); in the times–plus format, multiplication preceded addition ([e.g. a × b + c]). Spacing was either congruent or incongruent with standard operator precedence rules. In congruent trials, visual spacing grouped the multiplication operation more tightly (narrower spacing around "×"), consistent with its higher precedence (e.g. [a × b   +   c]). In incongruent trials, spacing conflicted with precedence by widening the space around multiplication (e.g. [a   ×   b + c]). Within each operation-order condition, half of the trials were congruent, and half were incongruent.

All participants completed the same set of 200 problems. Trial order was randomized independently for each participant using the default randomization procedure implemented in OpenSesame.

Each trial presented one arithmetic expression along with four response alternatives (see Table 1b). One alternative was the mathematically correct solution (e.g. for [6   +   3 × 4], [18]). A second alternative corresponded to the solution that would result if operator precedence were incorrectly reversed (e.g. [36]). The remaining two distractors were generated by adding or subtracting 2 from these two values. For example, for 6   +   3 × 4, potential distractors were [20] or [16] (±2 from [18]) and [38] or [34] (±2 from [36]). This procedure ensured matched parity across all response options. Across the experiment and within each experimental cell (Operation Order × Spacing Congruency), incorrect response magnitudes were balanced such that they were not systematically larger or smaller than either the correct or the *"operation-order error" alternative*. The overall distribution of stimulus types (Operation Order × Spacing

Congruency) was not significantly different: $\chi^2(3, n = 200) = 0.96, p = .811$. Similarly, the distribution of distractor answer types (±2 relative to correct vs. the *"operation-order error" alternative*) was balanced when addition appeared first, $\chi^2(3, n = 100) = 0.55, p = .907$, and when multiplication appeared first, $\chi^2(3, n = 100) = 0.55, p = .907$.

Symbols were displayed in black font (#000000, Droid Sans, size 40px) against a grey background (#dfdfdf). Narrowly spaced problems had a 64 px gap between operands, including the operation sign, while widely spaced problems had a 160 px gap (ratio 2:5). See Supplementary Materials, Figure S1, for an example stimulus, and Table S2 for a complete list of problems.

Math4Speed

The second task was M4S, aiming to assess participants' overall arithmetic fluency (Loenneker et al., 2025). While this speeded calculation task was originally designed for paper-and-pencil administration, here we used its computerized version, the same as in (Obidziński et al., 2025). Participants were instructed to solve the arithmetic problems mentally, as quickly as possible, without using external aids and without skipping any items. The task consisted of 200 items divided equally into four subtests, each covering a basic arithmetic operation. Items were presented in a fixed order across four blocks, as in the original M4S, with a time limit 120 seconds per operation type. The outcome measure was the total number of correctly solved items within the 2 minutes time limit per operation.

Math grades and demographic survey

The session ended with a short survey, in which participants were asked to give a rough estimate of the last grades attained in the subjects of mathematics and native language on the 0-9 scale (0-fail, 1–9 refer to grades when passed). They were asked about their gender, achieved education level and field of study.

Software

The procedure was programmed in the OpenSesame (v. 2.2.6) software (Mathôt et al., 2012) and presented online through the JATOS platform (v. 3.9.4) (Lange et al., 2015).

Preregistration, data availability and Ethics Committee Approval

The study was preregistered on Open Science Framework (https://doi.org/10.17605/OSF.IO/RG6DE). Data were collected between 03.02.2025 and 08.02.2025. The dataset, *R* code and materials are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/EZYSB)

The study was approved by the Rector's Committee for Research Ethics of the Jagiellonian University (1027.0041.1.2025).

Data filtering and deviations from the preregistration

We made the following deviations from the preregistration. We decided to apply a two-stage data trimming procedure. The first step was implemented immediately after data collection, when we noticed that some participants displayed inconsistent response patterns — specifically, a high proportion of incorrect answers accompanied by very short RTs. Therefore, in addition to excluding participants with accuracy below 25% (as preregistered), we increased the anticipation cutoff threshold from the preregistered 200 ms to 500 ms to exclude participants who had at least 15% of incorrect responses below this RT. As our task is quite complex, we concluded that this threshold could not be set too low. The value of 500 ms as a lower cutoff was chosen based on the analysis of the RTs' distribution and values reported in experiments on arithmetic problem-solving (Ronasi et al., 2018). This resulted in the exclusion of 17 participants from the 218 completed datasets, leaving a final sample of 201 participants.

All main analyses were conducted on the full sample of 201 participants. For the RTs analysis, additional data filtering was performed on the final dataset, which included trials from all 201 participants. We included only correct responses (87.5% of trials). A lower cutoff of 500 ms (revised from the preregistered 200 ms) and an upper cutoff of 3 *SD*s above the mean (as preregistered) were applied. After these exclusions, 82.9% of trials were retained. For comparison, using the 200 ms threshold would have retained 83% of trials; thus, the change did not substantially alter the inclusion criteria. Moreover, without RT trimming (considering correct trials only), the overall mean RT across participants was M = 5140 ms, SD = 5502 ms, and the mean RT for the fastest participant was M = 1664 ms, SD = 521 ms, further justifying the use of a 500 ms lower cutoff.

Data analysis

Data were analyzed using repeated measures analyses of variance (ANOVA) with two within-subject factors: Congruency (congruent vs. incongruent spacing) and Order (addition-first vs. multiplication-first). For each dependent measure (RT, accuracy and error rates), trial-level responses were first aggregated within each participant to obtain one summary value per condition (2 Congruency × 2 Order), and these participant-level scores were then entered into the repeated-measures ANOVAs. Thus, participants constituted the unit of analysis. Separate 2 × 2 ANOVAs were conducted for accuracy, RTs, and operation-order error rates. For the RT analysis, we included only correct responses, to obtain a clean estimate of processing speed under accurate computation, and to avoid confounding effects associated with error-related responses. Because repeated-measures ANOVA uses listwise deletion for observations with missing cells, the RT analysis was conducted only on participants with valid reaction times in all four experimental cells. Accuracy was calculated as the percentage of correctly solved problems. Mean accuracy and mean RTs were calculated at the participant level by averaging participants' condition-specific scores within each condition. For each condition (Congruency, Operation Order), the total number of errors and operation-order errors were calculated and reported descriptively. Error rates were calculated at the participant level within each condition and subsequently averaged across participants. In addition, the percentage of operation-order errors relative to all errors was computed to show how many of the total errors involved a violation of the precedence rule.

For the M4S task, each participant's total score was calculated as the number of correct responses completed within the 8-minute time limit, while separate scores were calculated for each operation type based on the number of correct responses within each 2-minute block. Overall mean scores and scores by operation type, along with their standard deviations (*SD*), were calculated. In addition, the mean error rate was computed as the number of errors divided by the total number of responses. Pearson correlation coefficients were computed between M4S scores and accuracy/RT measures from the Precedence and Proximity Task to assess relationships between arithmetic fluency and susceptibility to perceptual grouping effects. Similarly, we calculated Pearson correlation coefficients between participants' self-reported grades in mathematics and native language (as reported in the survey) and their M4S scores, as well as accuracy and RT measures from the main experimental task.

Drawing on Grice et al.'s (2020) approach, we conducted subsequent analyses calculating the percentage of participants for whom the differences in accuracy and RTs between in the congruent and incongruent conditions were consistent with the hypothesis. This approach aimed to investigate whether the effect observed at the group level is reflected by a majority of participants (Cipora et al., 2019).

In addition, we conducted a robustness analysis excluding participants whose accuracy fell below chance level (25%) in any individual condition. Because this exclusion criterion was not preregistered and does not affect the primary theoretical conclusions, we report the full-sample results as primary. This criterion identified 25 participants who showed near-zero (and considerably below-chance) performance in at least one cell, predominantly in the incongruent condition. The reduced subset included 176 participants. Results from this restricted sample are reported in addition to the main analysis to verify that the main effects were not driven by these extreme cases.

Results

Precedence and Proximity Task

Accuracy

Overall accuracy across 201 participants was 87.5% (M = 0.88, SD = 0.23).
Figure 1 presents raw accuracy as a function of Congruency and Operation Order.

Assumption checks

The assumptions of normality (Shapiro–Wilk tests in all conditions; all $p < .001$) and homogeneity (Levene's tests: $F(1, 802) = 115.54$, $p < .001$ for Congruency and

$F(1, 802) = 4.58$, $p = .033$ for Order) were violated. Given that accuracy showed ceiling effects (medians = 0.96), this violation was not unexpected. We used the [2*arcsin(sqrt(proportion_correct))] transformation to account for this (Hohol et al., 2024; Zar, 2010, p. 293). However, figures and descriptives, are presented in raw (untransformed) proportions of correct responses.
ANOVA results for untransformed accuracy are provided in the Supplementary Material 3.

Transformed accuracy

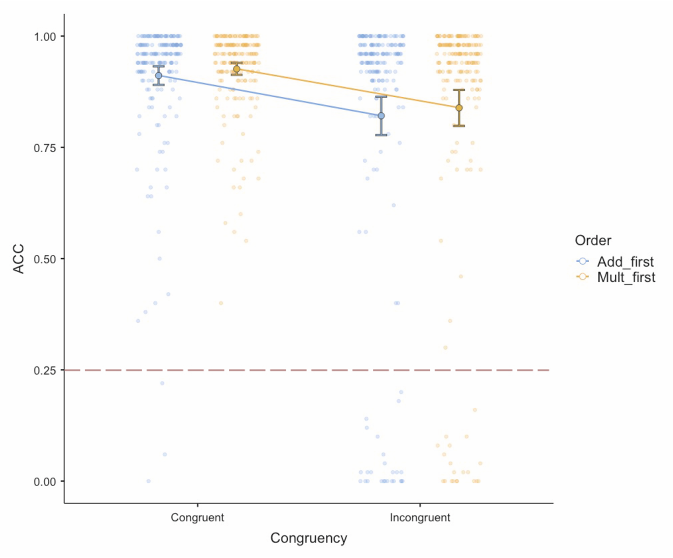
A 2 × 2 repeated-measures ANOVA was conducted to examine the effects of Congruency (congruent vs. incongruent) and Order (addition-first vs. multiplication-first) on performance scores. There was a significant main effect of Congruency, $F(1, 200) = 22.85$, $p < .001$, $\eta^2_p = .10$, indicating that participants performed better in the congruent spacing condition ($M = 0.92$, $SE = 0.01$) compared to the incongruent spacing condition ($M = 0.83$, $SE = 0.02$).

We did not find a significant effect of Order, $F(1, 200) = 2.75$, $p = .099$, $\eta^2_p = .01$, thus not justifying a robust difference in performance between the addition-first and multiplication-first conditions. We also did not find a significant interaction between Congruency and Order, $F(1, 200) = 0.95$, $p = .331$, $\eta^2_p = .01$, indicating that the effect of congruency on performance did not differ depending on Operation Order.

We calculated Pearson's correlation coefficients and found a moderate positive correlation between participants' self-reported mathematics grades (on a 0-9 scale) and overall accuracy ($r = .40$, $p < .001$) and a weak statistically significant correlation between grades in the native language and RTs ($r = .18$, $p = .013$). When accuracy was analyzed separately for the congruent and incongruent conditions, the Pearson correlation coefficients were as follows: mathematics grades – congruent: $r = .33$, $p < .001$, incongruent: $r = .36$, $p < .001$, native language grades – congruent: $r = .07$, $p = .298$), incongruent: $r = .19$, $p = .007$.

Additionally, the analysis of the proportion of participants for whom the difference in accuracy between incongruent and congruent stimuli was aligned with the theoretical predictions showed that this was the case for only 43.3% of the sample. A total of 23.1% had the same accuracy for both conditions. A total of 33.6% of participants performed more accurately on incongruent than on congruent stimuli. These results indicate substantial individual variability despite the significant group-level effect.

Figure 1
Untransformed accuracy scores by Congruency (congruent spacing vs. incongruent spacing) and Order (addition-first vs. multiplication-first).



Note. *Untransformed accuracy scores as a function of Congruency (congruent vs. incongruent spacing) and Order (addition-first vs. multiplication-first). The circles represent mean results per condition, error bars represent ±1 standard error of the mean. The dashed line represents the chance level (0.25). Accuracy was significantly higher in the congruent condition compared to the incongruent condition. The main effect of Order and the interaction between Congruency and Order were not significant.*

Robustness check without outliers

As shown in Fig. 1 [Untransformed accuracy scores by Congruency)], some participants who had overall accuracy above the chance (> 25%; i.e., met our inclusion criteria) displayed near-zero accuracy within individual conditions. The majority of these participants (21 out of 25) appeared to understand the precedence rule — as evidenced by above-chance performance in the congruent condition regardless, of Operation Order — yet performed below chance in the incongruent condition. Therefore, we retained them in the analysis. These participants' accuracy was 85.43% (*SE* = 3.25%) in the congruent condition and 4.14% (*SE* = 1.21%) in the incongruent condition.

To ensure that the overall effect was not driven by these participants, we conducted an additional analysis, in which we excluded participants who scored below 25% accuracy in any condition. After this exclusion, 176 participants remained in the dataset, and no effects on transformed accuracy remained significant. The main effect of Congruency was not significant, $F(1, 175) = 0.45$, $p = .50$, $\eta^2_p < .01$. The main effect of Order was also not significant, $F(1, 175) = 0.17$, $p = .68$, $\eta^2_p < .01$. We also did not find a significant interaction between Congruency and Order, $F(1, 175) = 0.67$, $p = .412$, $\eta^2_p < .01$.

Reaction Times analysis

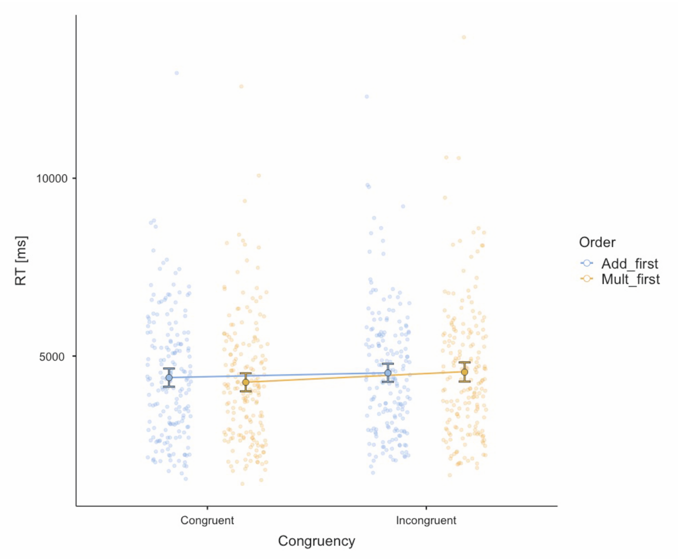
The assumption of normality (Shapiro-Wilk tests in all conditions, all $p < .001$) was violated, which is typical for RT data. The assumption of homogeneity of variance was met; Levene's tests indicated no significant differences in variances across levels of Congruency ($F(1, 779) = 0.02$, $p = .886$) or Order ($F(1, 779) = 0.00$, $p = .998$).

A 2 × 2 repeated-measures ANOVA was conducted to examine the effects of Congruency (congruent vs. incongruent) and Order (addition-first vs. multiplication-first) on RTs. Analysis included only participants with valid RT data in all four within-subject conditions; because RTs were calculated only for correct responses, missing data reduced the analyzable sample size to 186 participants. There was a significant main effect of Congruency, $F(1, 185) = 75.27$, $p < .001$, $\eta^2_p = .29$ indicating that participants responded faster in the congruent condition (M = 4331 ms, SE = 128 ms) compared to the incongruent condition (M = 4545 ms, SE = 132). We did not find a significant main effect of Order, $F(1, 185) = 2.95$, $p = .088$, $\eta^2_p = .02$. There were no significant differences in RTs between the addition-first and multiplication-first conditions. However, the interaction between Congruency and Order was significant, $F(1, 185) = 13.18$, $p < .001$, $\eta^2_p = .07$, indicating that the congruency effect on RTs was modulated by the Operation Order. For the multiplication-first condition, RTs were shorter for congruent problems (M = 4266 ms, SE = 129) compared to incongruent ones (M = 4561 ms, SE = 138 ms) with mean difference MD = −295 ms (SE = 30, $t(185) = -9.96$, $p < .001$, Cohen's d (paired-samples, d_z) = −0.730), whereas in the addition-first condition (congruent: M = 4397 ms, SE = 128 ms, incongruent: M = 4530 ms, SE = 128 ms), this difference was smaller (MD = −133 ms, SE = 37, $t(185) = -3.63$, $p = .002$, $d_z = -0.266$) [cf. Fig. 2].

We observed a weak negative correlation between participants' self-reported mathematics grades (on a 0-9 scale) and RTs ($r = -.39$, $p < .001$), as well as a weak but statistically significant correlation between grades in the native language and RTs ($r = -.28$, $p < .001$). When RTs were analyzed separately for the congruent and incongruent conditions, the Pearson correlation coefficients were as follows: mathematics grades – congruent: $r = -.39$, $p < .001$, incongruent: $r = -.37$, $p < .001$, native language grades – congruent: $r = -.25$, $p < .001$), incongruent: $r = -.30$, $p < .001$.

Analysis of how many participants showed a difference in RTs between incongruent and congruent stimuli was consistent with the theoretical predictions (similar to one conducted for accuracy) revealed that it was true for 73.4% of the sample. This percentage, as indicated by a two-sided binomial test, was significantly different from 50% ($p < .001$), showing a clear trend consistent with the predictions (faster RTs for congruent than for incongruent stimuli).

Figure 2
Reaction times (RTs) for correct responses across Congruency (congruent vs. incongruent) and Operation Order (addition-first vs. multiplication-first)



Note. *Reaction times (RTs) for correct responses as a function of Congruency (congruent vs. incongruent spacing) and Order (addition-first vs. multiplication-first). Error bars represent ±1 standard error of the mean. Participants responded significantly faster in the congruent condition than in the incongruent condition. A significant interaction indicated that the Congruency effect was more pronounced in the multiplication-first group, while RTs in the addition-first group showed little difference between congruent and incongruent problems. Our data showed no significant effect of Order.*

Robustness check without outliers

After excluding outliers who scored below 25% accuracy in any cell, the analysis on the remaining 176 participants showed a change only in the main effect of Order. A 2 × 2 repeated-measures ANOVA revealed a significant main effect of Congruency, $F(1, 175) = 126.85$, $p < .001$, $\eta^2_p = .42$, indicating that participants performed faster in congruent (M = 4240 ms, SE = 130 ms) than in incongruent (M = 4478 ms, SE = 135 ms) conditions. In contrast to the main analysis, there was also a significant main effect of Order, $F(1, 175) = 8.14$, $p = .005$, $\eta^2_p = .04$, suggesting that the order of operations affected performance (addition-first: M = 4391 ms, SE = 131, multiplication-first: M = 4326 ms, SE = 135 ms). Importantly, the Congruency × Order interaction was significant, $F(1, 175) = 8.77$, $p = .004$ $\eta^2_p = .05$, indicating that the effect of Congruency depended on the Operation Order. Post hoc tests showed that for the multiplication-first condition, RTs were significantly shorter for congruent problems (M = 4177 ms, SE = 132 ms) than for incongruent ones (M = 4476 ms, SE = 140), with mean difference MD = −299 ms, SE = 29 ms ($t(175) = -10.47$, $p < .001$, $d_z = -0.789$). In the addition-first condition, RTs were also shorter for congruent (M = 4303 ms, SE = 130) than incongruent problems (M = 4480 ms, SE = 132 ms), with mean difference MD = −177 ms (SE = 30 ms, $t(175) = -5.84$, $p < .001$, $d_z = -0.440$). For multiplication-first condition, RTs were shorter for congruent problems compared to incongruent ones, similarly to the main analysis.

Additionally, Table 2 presents summarized descriptive statistics (means and standard deviations) for accuracy and RT across Congruency and Operation Order conditions in Precedence and Proximity task.

Table 2 Descriptive Statistics for Precedence and Proximity Task

	Accuracy (M±SD) [%]		**RT (M±SD) [ms]**	
	Congruent Spacing [%]	Incongruent Spacing [%]	Congruent Spacing [ms]	Incongruent Spacing [ms]
Addition first	91.16±15.2	82.1 ± 31.0	4516 ± 1797	4530 ± 1760
Multiplication first	92.7 ±9.5	83.9 ± 28.6	4357 ± 1791	4553 ±1875

Note. Means (M) and standard deviations (SD) were calculated across participants separately for each condition defined by operation (addition vs. multiplication) and spacing (congruent vs. incongruent). Accuracy represents the proportion of correct responses. Reaction times (RTs) were computed based on correct trials only. All statistics were calculated using available observations with missing values removed. Accuracy descriptives were based on n = 201, and RT descriptives were based on n = 186.

Error type analysis

Error rates (total number of errors / number of trials per condition) and Operation-Order Error Rates (total number of operation-order errors / number of trials per condition) are presented in Table 3.

Table 3

Error Rates and Operation-Order Error Rates by Congruency and First Operation Type

Congruency	**First Operator (Order)**	**Total number of errors**	**Total number of operation-order errors**	**Total error rate**	**Operation-order error rate**	**Operation-order errors as % of total error number**
congruent	+	888	319	8,8%	3.2%	19.8%
congruent	×	738	133	7,3%	1.3%	13.5%
incongruent	+	1798	1237	17,9%	12.3%	32.5%
incongruent	×	1615	1048	16,1%	10.4%	29.9%

Note. Values represent the total number of errors and corresponding error rates (overall errors and operation-order errors) across conditions (Congruency: congruent vs. incongruent; Order: addition-first vs. multiplication-first). Aggregated absolute error counts are presented descriptively. Error rates were computed for each participant within each condition and then averaged across participants. The proportion of operation-order errors relative to total errors may differ from aggregated estimates due to variation in the number of errors across participants. All ANOVAs were conducted at the participant level, using per-participant means within each condition.

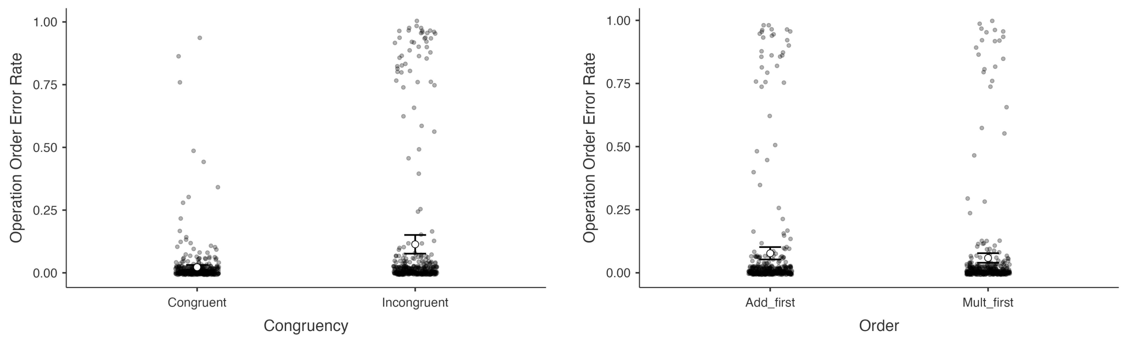
A 2 × 2 repeated-measures ANOVA was conducted to examine the effects of Congruency (congruent vs. incongruent) and Order (addition-first vs. multiplication-first)

on the operation-order error rates. There was a significant main effect of Congruency, $F(1, 200) = 25.74$, $p < .001$, $\eta^2_p = .11$ with participants with lower operation-order error rates in the congruent condition ($M = 0.022$, $SE = 0.005$) compared to the incongruent condition ($M = 0.114$, $SE = 0.019$). The main effect of Order was also significant, $F(1, 200) = 5.54$, $p = .020$, $\eta^2_p = .03$, indicating that participants had higher operation-order error rate when performing addition-first ($M = 0.077$, $SE = 0.012$) than when performing multiplication-first ($M = 0.059$, $SE = 0.009$). The interaction between Congruency and Order was not significant, $F(1, 200) = 0.01$, $p = .912$, $\eta^2_p = .00$, showing that in our sample the effect of congruency on operation-order error rates did not vary by operation order.
However, operation-order error rates were generally low across conditions (in absolute counts for Congruency: congruent condition ($M = 1.12$, $SE = 0.23$), incongruent condition ($M = 5.68$, $SE = 0.94$) and for Order: addition-first ($M = 3.87$, $SE = 0.62$), multiplication-first ($M = 2.94$, $SE = 0.48$)), indicating a floor effect.

To examine whether these effects reflected general differences in task difficulty, an additional 2 × 2 repeated-measures ANOVA was conducted on total error rates. This analysis similarly revealed a significant main effect of Congruency, $F(1, 200) = 23.96$, $p < .001$, $\eta^2_p = .11$, indicating higher overall error rates in incongruent than congruent trials (congruent: $M = 0.081$, $SE = 0.008$; incongruent: $M = 0.170$, $SE = 0.018$). The main effect of Order was also significant, $F(1, 200) = 4.04$, $p = .046$, $\eta^2_p = .02$, indicating higher error rates in the addition-first condition ($M = 0.097$, $SE = 0.012$) than in the multiplication-first condition ($M = 0.085$, $SE = 0.009$). The interaction between Congruency and Order was not significant, $F(1, 200) = 0.66$, $p = .418$, $\eta^2_p = .003$, mirroring the pattern observed for operation-order errors.

Figure 3

Distribution of Operation-Order Error Rates by Congruency (left) and Operation Order (right).



Note. Distribution of Operation-Order Error Rates by Congruency (left panel) and Operation Order (right panel). Empty circles represent mean Operation-Order Rate per condition, error bars represent ±1 standard error of the mean. Participants made significantly fewer operation order errors in the congruent condition compared to the incongruent condition. Additionally, error rates were higher in the addition-first group than in the multiplication-first group.

Robustness check without outliers

After excluding outliers who scored below 25% accuracy per cell, the main effect of Order was no longer significant, $F(1, 175) = 1.37$, $p = .24$, $\eta^2_p = .01$. The main effect of Congruency remained significant $F(1, 175) = 11.37$, $p < .001$, $\eta^2_p = .06$ with higher operation-order error rate for incongruent condition ($M = 0.016$, $SE = 0.002$) in comparison to congruent condition ($M = 0.012$, $SE = 0.002$). The interaction between Congruency and Order remained not significant, $F(1, 175) = 0.04$, $p = .85$, $\eta^2_p < .01$.

Descriptively, at the participant level, among the participants identified as outliers who demonstrated an understanding of the precedence rule (n = 21), mean total error rates were substantially higher in the incongruent condition (addition-first: 96.6%, multiplication-first: 95.1%) than in the congruent condition (15.6% and 13.5%, respectively). Mean operation-order error rates showed the same pattern, with higher rates in incongruent trials (addition-first: 86.3%, multiplication-first: 84.5%) compared to congruent trials (4.8% and 3.5%, respectively).

Math4Speed

The M4S score reflected the number of correct responses per 2-minute operation block, with a total of 8 minutes to complete all four operations. The overall mean M4S score (mean number of correct answers in given time) was 79.2 ($SD = 33.0$), the lowest score was 28, and the highest 181. On average, participants achieved the highest score in the addition subtest ($M = 21.9$, $SD = 8.3$) and the lowest in subtraction ($M = 16.2$, $SD = 7.2$). Error rates were lowest for addition ($M = 0.07$, $SD = 0.09$) and highest for division ($M = 0.12$, $SD = 0.14$). Table 4 provides a detailed summary of the scores across different operation types.

Table 4
Descriptive Statistics for Math4Speed Scores by Operation Type

Operation type	**Mean Correct**	***SD* Correct**	**Mean Error Rate**	***SD* Error Rate**
Addition	21.9	8.2	0.066	0.086
Subtraction	16.2	7.2	0.109	0.141
Multiplication	20.1	10.2	0.110	0.124
Division	20.9	11.2	0.116	0.135

Note. Values represent number of correct answers and error rates (M, SD) for each arithmetic operation (Addition, Subtraction, Multiplication, Division).

Additionally, we observed a moderate positive correlation between participants' self-reported mathematics grades (on a 0-9 scale) and M4S overall scores ($r = .44$,

$p < .001$). In contrast, we did not find a statistically significant correlation between grades in the native language and M4S scores ($r = .10$, $p = .14$).

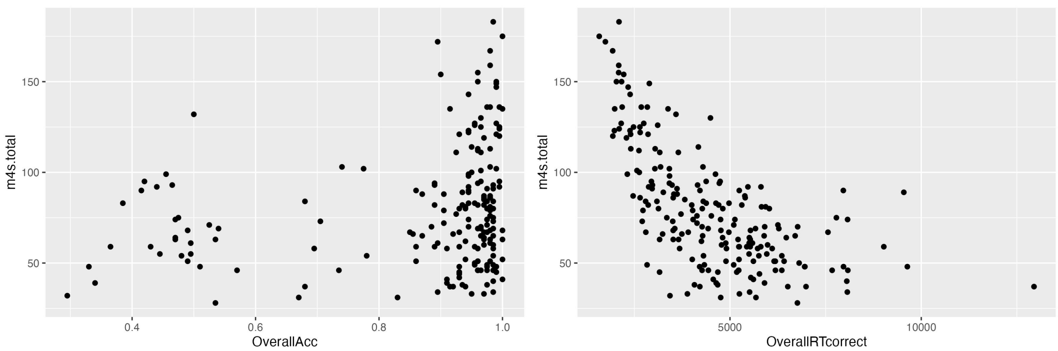
Descriptive results for outliers

For the outliers (n = 25), the overall mean M4S score was 65.6 (*SD* = 23.3), with mean scores of 18.8 ± 6.2 for addition, 13.2 ± 5.7 for subtraction, 17.4 ± 7.4 for multiplication, and 16.2 ± 8.7 for division.

Precedence and Proximity Task and M4S

Pearson correlations between M4S scores and results from the Precedence and Proximity task were calculated. We observed a weak positive correlation between participants' M4S scores and their overall accuracy in the Precedence and Proximity Task ($r = .21$, $p = .002$). This suggests that individuals who performed better on the M4S task also tended to be more accurate in the Precedence and Proximity Task. In contrast, there was a moderate, statistically significant negative correlation between M4S scores and RTs for correct responses in the Precedence and Proximity Task ($r = –.64$, $p < .001$). This indicates that participants with higher M4S scores tended to respond faster in the Precedence and Proximity Task. Additional correlations are reported in Supplementary Material 4.

Figure 4
Pearson correlations between M4S scores and results from the Precedence and Proximity task



Note. Scatterplots illustrating the relationship between M4S scores and performance accuracy (left), and between M4S scores and RTs for correct answers (right).

Robustness check without outliers

After excluding outliers, the correlation pattern did not change substantially: between M4S and overall accuracy in the Precedence and Proximity Task, $r = .16$, $p = .04$, and between M4S and RTs for correct trials, $r = – .65$, $p < .001$.

Discussion

Overview

This study aimed to investigate whether arithmetic problem-solving is affected by perceptual grouping biases — specifically, visual factors such as spacing and proximity between operands presented in a forced-choice format. Whereas prior adult studies primarily relied on production-based response formats and often relatively small samples, the present study tested whether spacing effects generalize to a forced-choice paradigm with more naturalistic spacing manipulations and sufficient power to detect small within-subject effects. Our results are generally consistent with Landy and Goldstone's (2010) findings and subsequent replications (Braithwaite et al., 2016; Harrison et al., 2020; Jiang et al., 2014; Rivera & Garrigan, 2016), corroborating that visual layout can affect mathematical performance. By employing a forced-choice format, we also introduced a novel way of testing spatial biases in arithmetic problem-solving within an online environment, which allowed more precise measurement of RTs.

Validation of hypotheses

Consistent with H1, congruent spacing was associated with higher accuracy, shorter RTs, and more frequent correct application of operation order than incongruent spacing. This supports the view that perceptual grouping can facilitate arithmetic processing when visual cues align with formal precedence rules. However, an additional robustness check showed that the congruency effect on accuracy in our sample may have been amplified by a relatively small subgroup of participants (n = 25) with markedly lower accuracy in the incongruent condition. Despite being reminded of the precedence rule (and presumably having learned it during prior education), they appeared to base their calculations consistently on the physical proximity of numbers rather than on the conventional order of operations. After excluding these individuals, in robustness analysis conducted on the remaining 176 participants, no effects on transformed accuracy remained significant. Thus, the observed effect may, at least to some extent, reflect a highly consistent pattern in some individuals rather than interference-based lower accuracy across most participants, who otherwise perform the task accurately in all conditions. This may be theoretically important, because it suggests that the congruency effect on accuracy may reflect at least two partially distinct phenomena: a broad, relatively mild processing cost visible across many participants, and a much stronger susceptibility to misleading visual grouping in a smaller subgroup.

The RT findings provided support for H1. RTs were shorter in congruent than incongruent spacing conditions for most participants, indicating reduced cognitive load when perceptual cues aligned with the expected operation order. Importantly, unlike the effects on accuracy, this was not driven mostly by individuals unable to perform incongruent trials correctly. Additional analysis yielded that 73.4% of participants showed faster RTs for congruent than incongruent stimuli, a proportion significantly above chance, and aligned with theoretical predictions. We interpret this as a typical pattern, where cognitive costs, though not always reflected in accuracy, manifest in longer RTs.

Similarly, operation-order errors were more frequent in incongruent than congruent conditions, indicating that misleading spacing can bias not only overall performance but also the order in which operations are carried out. This further supports the claim that incongruent spacing is not merely associated with general performance declines but actively misguides participants, leading them to perform operations in the wrong order. However, given the overall low rate of operation-order errors, a floor effect limits the strength of this conclusion. A robustness check without outliers showed that the main effect of Congruency remained significant, but operation-order error rates became notably lower, again suggesting that outliers affected the results.

Support for H2 was weaker. We did not find a significant main effect of Operation Order on accuracy or RTs. However, the interaction between Congruency and Operation Order was significant for RTs, indicating that responses were faster only for multiplication-first expressions with congruent spacing. Robustness analyses excluding outliers revealed the significant main effect of Order with slightly shorter RTs for multiplication-first condition. We expected such results, as the order of problem-solving aligns with the reading direction in Western cultures. Analysis of errors showed that participants in general made more operation-order errors in addition-first than in multiplication-first problems, but the main effect of Order was again driven by the outlier group, disappearing once they were excluded from the analysis. Moreover, error rates should be interpreted with caution due to a floor effect. More broadly, this suggests that while operation order may modulate performance under some circumstances, it was not the primary factor driving participants' behavior in the current task.

Compared to prior research, some congruency effects in the present study were weaker (e.g., in accuracy), and some effects observed in the original Landy and Goldstone adult paradigm were not replicated, specifically the main effect of Operation Order in RTs for the whole sample. Notably, later studies such as Rivera and Garrigan (2016) and Jiang et al. (2014) confirmed that spatial manipulations can increase errors and bias solution procedures, but they did not test exactly the same pattern of order-related effects. Therefore, the current design does not allow us to determine whether the differences observed here are attributable, for example, to the more naturalistic spacing manipulation in our design or to the forced-choice response format.

Finally, in line with H3, participants with higher mathematical fluency (as measured by M4S scores) were more accurate and significantly faster in solving problems on the Precedence and Proximity Task. This pattern suggests that basic calculation fluency facilitates efficient processing in the Precedence and Proximity Task, although it does not fully explain susceptibility to spacing effects. Note that the M4S task is, by design, a speeded calculation test that favors individuals who solve arithmetic problems both quickly and accurately. By contrast, precise and careful problem-solvers who require more time tend to score lower on the M4S. Additionally, the M4S task comprises calculations of only one operation per block and does not demand the same level of cognitive control as a more complex task that includes mixed expressions and requires maintaining the proper order of operations. This may account for the moderate correlation with RTs and the weak correlation with accuracy in the Precedence and Proximity Task. Importantly, the results after excluding outliers did not change substantially, which may indirectly suggest that the main difference between outliers and the rest of the group did not lie solely in arithmetic fluency and competence.

Taken together, the present findings support the conclusion that arithmetic performance is shaped not only by formal knowledge and calculation skill, but also by the perceptual organization of symbolic notation.

Perceptual biases and arithmetic fluency

The question arises as to whether lower mathematical competencies are linked to increased susceptibility to errors resulting from perceptual grouping. Higher mathematical competence and stronger executive control may facilitate the ability to ignore irrelevant perceptual cues, or individual susceptibility to Gestalt principles may increase vulnerability to errors. In Harrison et al.'s (2020) research students with higher prior performance made fewer errors, suggesting a link between conceptual understanding and computational accuracy, although incongruent spacing still impaired performance, even after controlling for prior ability. However, in studies with children (Braithwaite et al., 2016; Harrison et al., 2020), adaptive methods were used to control for prior mathematical performance — specifically, the Math Garden algorithm, which selects problems with an estimated 75% success rate, and prior performance data from the ASSISTments online tutoring system. For the first time, we applied an independent measure of arithmetic fluency to examine its relationship with perceptual biases in arithmetic problem-solving. We assessed participants' arithmetic fluency directly rather than controlling for their skill level. Participants with higher M4S scores were more accurate (weak positive correlation) and significantly faster (moderate negative correlation) in solving problems on the Precedence and Proximity Task.

In our data, M4S overall scores were moderately correlated with participants' mathematics grades, but not with grades from the native language. The lack of correlation with native language grades provides evidence for the discriminant validity of the M4S measure, suggesting that it captures domain-specific mathematical abilities rather than broader academic performance. Similarly, we found a moderate correlation between accuracy and RTs in Precedence and Proximity Task and grades in math, and only weak correlation with the language grades (although the correlations with native language grades were larger for incongruent RTs and incongruent accuracy). These results suggest a consistent link between overall mathematical competence and performance in the Precedence and Proximity Task, as reflected in both accuracy and response times.

Outliers and individual susceptibility to visual cues

Arguably, while perceptual grouping influences the processing of mathematical expressions in the general population, some individuals seem to be more susceptible to the visual properties of problem presentation. In particular, we observed a pronounced impact on the reported effects — especially regarding Congruency — within the outlier group, consisting of participants whose overall accuracy was above chance (> 25%) but who nevertheless showed near-zero performance in at least one condition. As most outliers (21 out of 25) demonstrated an understanding of the precedence rule — evidenced by above-chance performance in the addition-first trials — we decided not to exclude them from the main analysis. They displayed a striking performance pattern — high accuracy in congruent conditions, but extremely high error rates in incongruent ones. Even though the outlier group had lower M4S mean scores (66 ± 23) compared with overall mean score (79 ± 33), they demonstrated that they possess basic arithmetic skills. Their responses suggest a strong influence of perceptual grouping, as if they interpreted closely grouped operations as being enclosed in brackets. This behavior may indicate heightened susceptibility to misleading visual cues, despite a general understanding of the precedence rule. However, it is also possible that some participants misunderstood the instructions or were simply unable to complete the task correctly. A similar case, involving a subgroup with null results on basic cognitive tasks such as the Stop-Signal Task and the Go/No-Go Task, was reported by Chuderski et al. (2012) and classified by the authors as genuine cases of goal neglect. Jiang et al. (2014) also noted that the spatial features of problem presentation influenced some participants more than others, as error frequency varied considerably across individuals. Based on developmental research by Braithwaite et al. (2016) we cannot rule out the possibility that our outlier group followed an opportunistic selection strategy, solving spatially grouped items first, which may have led to errors. However, our study design and relatively small number of outliers do not allow for a more detailed interpretation of this behavior.

Theoretical and Educational Implications of Spacing Effects

Using more naturalistic stimuli than those used by Landy & Goldstone (2010), we demonstrated that subtle manipulations of the graphical layout can facilitate arithmetic problem-solving and may serve as a tool in everyday educational settings. During the stage of acquiring abstract rules, visual cues can support the construction of basic mathematical concepts. At later learning stages involving practice, application, or evaluation of knowledge and skills, task difficulty can be increased by employing uniform spacing or even incongruent formats. Such approaches may help assess a deeper understanding of abstract principles and foster more focused attentional control. This aligns with the idea that mathematical reasoning prioritizes precision over approximation and emphasizes accuracy and consistency over heuristics or everyday habits.

These findings demonstrate that spatial features of stimulus presentation influence calculation efficiency in terms of both accuracy and computation speed. Regarding the interplay between Congruency and Operation Order, our conceptual replication corroborates Landy and Goldstone’s (2010) findings as well as later follow ups with adults (Jiang et al., 2014; Rivera & Garrigan, 2016). In these studies, participants showed higher accuracy and faster RTs in congruent conditions and made more precedence errors under incongruent spacing. However, in our study, the Operation Order effect was either small (precedence error rates) or not statistically significant (accuracy and RTs). Prior adult studies, especially (Landy & Goldstone, 2010), reported evidence that expressions beginning with addition (e.g., plus–times) are often more difficult in terms of accuracy and RT because they conflict more strongly with left-to-right reading habits and precedence structure. In our sample and task format, spacing congruency appeared to be more robust determinant of performance than operation order. Moreover, the absence of a significant interaction between Congruency and Operation Order for precedence error rate suggests that spacing affected precedence errors similarly across operation orders, rather than more strongly in one expression type as reported in the replicated Experiment 3. In our replication, we tested more than 200 participants — likely making this the first adult study in this area to achieve sufficient statistical power (>.80) to detect a within-subject effect size of Cohen's $d = 0.2$ and a Pearson correlation of $r = .2$. This level of power should yield more precise effect size estimation (see Table S1 in the Supplementary Materials for number of participants details in previous studies on adults). We demonstrated that the described perceptual bias, involving differences in proximity between operands in solving arithmetical expressions, persists in the forced-choice format. In the production paradigm used in the original experiment by Landy and Goldstone (2010), as well as in subsequent replications (Braithwaite et al., 2016; Harrison et al., 2020; Jiang et al., 2014; Rivera & Garrigan, 2016), RT was measured as the interval between stimulus presentation and the first key press. By contrast, the multiple-choice format captures the full duration of problem processing until the final decision.

Taken together with earlier work, the present findings support the view that arithmetic notation is processed not only as an abstract symbolic system, but also as a visual layout that guides attention and decision-making. Gestalt principles that organize human perception may interfere with symbolic reasoning and disrupt rule-based calculations. Solving problems presented in incongruent formats requires selective attention and the inhibition of misleading perceptual interpretations. Many authors argue that flexibility is a key feature in the acquisition and application of mathematical skills (Cipora, He, et al., 2020; Greer, 2009; Hong et al., 2023). The ability to choose among various strategies and to switch between abstract and concrete representations is fundamental to developing mathematical competence. From an educational perspective, the findings suggest that spacing and visual layout may function as subtle perceptual scaffolds during the acquisition of order-of-operations rules. An individual's capacity to alternate between suppressing perceptual cues when they mislead and relying on them when they support correct calculations may be a crucial factor when perceptual biases come into play.

Similar research on the effects of proximity on processing numerical information has been conducted not only in symbolic tasks, as mentioned above, but also in non-symbolic numerosity processing. Experiments on the groupitizing phenomenon

(Maldonado Moscoso et al., 2025) have shown that objects presented in groups or clusters are enumerated more quickly. Moreover, the precision of estimating clustered items decreases when a concurrent simple calculation task is performed, whereas estimation accuracy for ungrouped items remains unaffected. These findings align with our results, as we observed shorter RTs, higher accuracy, and lower error rates in congruent conditions, where terms in equations were closer together. Participants behaved as if they interpreted closely grouped expressions as distinct units or even as bracketed elements.

Together with earlier results, our study provides another piece of evidence for the numerous links between numerical and spatial processing, especially in cases where such links are not beneficial for solving the task. Moreover, when considered alongside findings from various paradigms, it contributes to a broader picture of space and numbers being closely intertwined in human cognition (Cipora et al., 2018).

Limitations and future research directions

Given the correlational design of our study, we cannot determine to what extent the observed spacing effect reflects predominantly perceptual grouping biases or factors strengthened through educational practice. In mathematical training, visual cues such as parentheses or subtle spacing adjustments in printed materials are commonly used. A further limitation is that the present design does not allow us to disentangle low-level perceptual grouping effects from overlearned notation-specific conventions acquired through mathematical instruction, as both may contribute to the observed congruency advantage. Nevertheless, we argue that because perceptual grouping is a well-documented and extensively studied phenomenon (Brooks, 2015; Wagemans et al., 2012), it should be recognized as a significant factor and may serve as an effective tool for teaching abstract rules. At the stage of assessing understanding of the order-of-operations rule, problems can be presented in congruent, incongruent, or evenly spaced formats to test whether individuals grasp the abstract rule and can apply it despite layout difficulty. This approach also helps identify those more susceptible to perceptual grouping errors, enabling more targeted interventions (Ottmar et al., 2015).

It would be valuable to further investigate the subgroup of participants who showed a striking pattern of high accuracy in congruent conditions but extremely high error rates in incongruent ones. In our study, the sample of outliers was too small to draw reliable conclusions. However, considering that individuals with similar behavioral characteristics have been noted in previous research (Jiang et al., 2014), future studies could examine whether this pattern recurs across different populations. Research should focus on the individual prevalence of the perceptual grouping effect in arithmetic, as some individuals appear more error-prone under incongruent spacing.

The next avenue for future studies is to explore how perceptual cues could be used in teaching the early stages of mathematics. Previous research shows that kindergarten children typically lack groupitizing skills, which begin to develop in Grades 1–3 (Starkey &

McCandliss, 2014). Braithwaite et al. (2016) found that in children aged 7–12, susceptibility to narrow spacing and the perceptual grouping effect increased with grade level. By contrast, Harrison et al. (2020) reported that older students (Grades 5–12, approximately 10-18 years old) made more errors and showed reduced performance when solving expressions with incongruent spacing, suggesting that perceptual grouping, rather than age or skill, was the key factor influencing problem-solving accuracy. It is worth investigating whether adults with lower mathematical skills are particularly susceptible to visual cues, and how mathematical competencies relate to susceptibility to perceptual grouping bias.

Conclusions

This study replicated the key findings of Landy and Goldstone (2010), confirming that visual layout significantly influences arithmetic performance. Our replication was successful despite the use of a different response format (forced-choice vs. production), demonstrating the overall robustness of the spacing effect on mental calculations. The forced-choice format allowed RT to reflect the time until the final decision rather than merely response onset, and importantly, the congruency effect remained evident. Our results demonstrate that visuospatial influences on arithmetic reasoning persist across task formats. Participants were more accurate and faster in congruent spacing conditions, where perceptual grouping aligned with conventional operation order. In contrast, incongruent spacing increased operation-order errors — particularly when problems began with addition — indicating that misleading visual cues can disrupt the correct application of the precedence rule. The broader implication is that symbolic reasoning remains partly constrained by the visual organization of notation. Furthermore, higher mathematical fluency (as measured by M4S scores) was associated with better performance and faster responses, though the correlation with accuracy was weak. Some individuals appear to be more susceptible to visual cues in mathematical expressions, which affects their performance especially in incongruent conditions. It is worth examining whether a similar group of outliers emerges in other research involving perceptual biases.

References

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The Future of Embodied Design for Mathematics Teaching and Learning. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.00147

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L. J., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition*, *149*, 40–55. https://doi.org/10.1016/j.cognition.2016.01.004

Brooks, J. L. (2015). Traditional and new principles of perceptual grouping. In J. Wagemans (Ed.), *The Oxford Handbook of Perceptual Organization* (pp. 57–87). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199686858.013.060

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1831–1840. https://doi.org/10.1098/rstb.2009.0028

Campbell, J. I. D. (2015). How Abstract is Arithmetic? In R. Cohen Kadosh & A. Dowker (Eds), *The Oxford Handbook of Numerical Cognition*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.024

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence but executive control does not. *Intelligence*, *40*(3), 278–295. https://doi.org/10.1016/j.intell.2012.02.010

Cipora, K., Haman, M., Domahs, F., & Nuerk, H.-C. (2020). Editorial: On the Development of Space-Number Relations: Linguistic and Cognitive Determinants, Influences, and Associations. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00182

Cipora, K., He, Y., & Nuerk, H.-C. (2020). The spatial–numerical association of response codes effect and math skills: Why related? *Annals of the New York Academy of Sciences*, *1477*(1), 5–19. https://doi.org/10.1111/nyas.14355

Cipora, K., Patro, K., & Nuerk, H.-C. (2015). Are spatial-numerical associations a cornerstone for arithmetic learning? The lack of genuine correlations suggests no. *Mind, Brain, and Education*, *9*(4), 190–206. https://doi.org/10.1111/mbe.12093

Cipora, K., Schroeder, P. A., Soltanlou, M., & Nuerk, H.-C. (2018). More Space, Better Mathematics: Is Space a Powerful Tool or a Cornerstone for Understanding Arithmetic? In K. S. Mix & M. T. Battista (Eds), *Visualizing Mathematics: The Role of Spatial Reasoning in Mathematical Thought* (pp. 77–116). Springer International Publishing. https://doi.org/10.1007/978-3-319-98767-5_4

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Goebel, S., Willmes, K., Pesenti, M., Schiltz, C., & Nuerk, H.-C. (2019). *A Minority pulls the sample mean: On the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect*. OSF. https://doi.org/10.31234/osf.io/bwyr3

Decker-Woodrow, L. E., Mason, C. A., Lee, J.-E., Chan, J. Y.-C., Sales, A., Liu, A., & Tu, S. (2023). The Impacts of Three Educational Technologies on Algebraic Understanding in the Context of COVID-19. *AERA Open*, *9*, 23328584231165919. https://doi.org/10.1177/23328584231165919

Ding, Y., Wang, Q., Liu, R.-D., Trimm, J., Wang, J., Feng, S., Hong, W., & Yang, X.-T. (2024). Working Memory Load, Automaticity, and Problem Solving in College Engineering Students: Two Applications. *SAGE Open*, *14*(4), 21582440241305082. https://doi.org/10.1177/21582440241305082

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The Education of Perception. *Topics in Cognitive Science*, *2*(2), 265–284. https://doi.org/10.1111/j.1756-8765.2009.01055.x

Greer, B. (2009). Representational flexibility and mathematical expertise. *ZDM*, *41*(5), 697–702. https://doi.org/10.1007/s11858-009-0211-7

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455. https://doi.org/10.1177/2515245920922982

Harrison, A., Smith, H., Hulse, T., & Ottmar, E. R. (2020). Spacing Out! Manipulating Spatial Features in Mathematical Expressions Affects Performance. *Journal of Numerical Cognition*, *6*(2), 186–203. https://doi.org/10.5964/jnc.v6i2.243

Hawes, Z., & Ansari, D. (2020). What explains the relationship between spatial and mathematical skills? A review of evidence from brain and behavior. *Psychonomic Bulletin & Review*, *27*(3), 465–482. https://doi.org/10.3758/s13423-019-01694-7

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, *10*(4), 389–395. https://doi.org/10.3758/BF03202431

Hohol, M., Cipora, K., Willmes, K., & Nuerk, H.-C. (2017). Bringing Back the Balance: Domain-General Processes Are Also Important in Numerical Cognition. *Frontiers in Psychology*, *8*, 499. https://doi.org/10.3389/fpsyg.2017.00499

Hohol, M., Szymanek, P., & Cipora, K. (2024). Analogue magnitude representation of angles and its relation to geometric expertise. *Scientific Reports*, *14*(1), 8997. https://doi.org/10.1038/s41598-024-59521-6

Hong, W., Star, J. R., Liu, R.-D., Jiang, R., & Fu, X. (2023). A Systematic Review of Mathematical Flexibility: Concepts, Measurements, and Related Research. *Educational Psychology Review*, *35*(4), 104. https://doi.org/10.1007/s10648-023-09825-2

Jiang, M. J., Cooper, J. L., & Alibali, M. W. (2014). Spatial factors influence arithmetic performance: The case of the minus sign. *Quarterly Journal of Experimental Psychology (2006)*, *67*(8), 1626–1642. https://doi.org/10.1080/17470218.2014.898669

Kelley, T. A., & Yantis, S. (2009). Learning to attend: Effects of practice on information selection. *Journal of Vision*, *9*(7), 16. https://doi.org/10.1167/9.7.16

Kirshner, D. (1989). The Visual Syntax of Algebra. *Journal for Research in Mathematics Education*, *20*(3), 274–287. https://doi.org/10.2307/749516

Landy, D., Allen, C., & Zednik, C. (2014). A perceptual account of symbolic reasoning. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00275

Landy, D., & Goldstone, R. L. (2007a). Formal notations are diagrams: Evidence from a production task. *Memory & Cognition*, *35*(8), 2033–2040. https://doi.org/10.3758/BF03192935

Landy, D., & Goldstone, R. L. (2007b). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 720–733. https://doi.org/10.1037/0278-7393.33.4.720

Landy, D., & Goldstone, R. L. (2010). Proximity and Precedence in Arithmetic. *Quarterly Journal of Experimental Psychology*, *63*(10), 1953–1968. https://doi.org/10.1080/17470211003787619

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Loenneker, H. D., Cipora, K., Artemenko, C., Soltanlou, M., Bellon, E., De Smedt, B., García-Orza, J., Giannouli, V., Gutiérrez-Cordero, I., Lipowska, K., van Dijck, J.-P., Yao, X., Nuerk, H.-C., & Huber, J. F. (2025). Math4Speed: A freely available measure of arithmetic fluency. *Canadian Journal of Experimental Psychology-Revue Canadienne de Psychologie Experimentale*, *79*(2), 212–220. https://doi.org/10.1037/cep0000347

Lonnemann, J., Krinzinger, H., Knops, A., & Willmes, K. (2008). Spatial representations of numbers in children and their connection with calculation abilities. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *44*(4), 420–428. https://doi.org/10.1016/j.cortex.2007.08.015

Maldonado Moscoso, P. A., Anobile, G., Maduli, G., Arrighi, R., & Castaldi, E. (2025). "Groupitizing": A Visuo-Spatial and Arithmetic Phenomenon. *Open Mind*, *9*, 121–137. https://doi.org/10.1162/opmi_a_00181

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Obidziński, M., Bażela, N., & Hohol, M. (2025). More gist, better math: Fuzzy-trace theory-based investigation of the relationship between long-term memory and mathematical skills. *Cognition*, *263*, 106212. https://doi.org/10.1016/j.cognition.2025.106212

Ottmar, E., Landy, D., Goldstone, R. L., & Weitnauer, E. (2015). Getting From Here to There! : Testing the Effectiveness of an Interactive Mathematics Intervention Embedding Perceptual Learning. 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 Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*. cognitivesciencesociety.org. https://mindmodeling.org/cogsci2015/papers/0311/index.html

Poincaré, H. (1952). *Science and hypothesis*. Dover Publications.

Rivera, J., & Garrigan, P. (2016). Persistent perceptual grouping effects in the evaluation of simple arithmetic expressions. *Memory & Cognition*, *44*(5), 750–761. https://doi.org/10.3758/s13421-016-0593-z

Ronasi, G., Fischer, M. H., & Zimmermann, M. (2018). Language and Arithmetic: A Failure to Find Cross Cognitive Domain Semantic Priming Between Exception Phrases and Subtraction or Addition. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01524

Shaki, S., Fischer, M. H., & Göbel, S. M. (2012). Direction counts: A comparative study of spatially directional counting biases in cultures with different reading directions. *Journal of Experimental Child Psychology*, *112*(2), 275–281. https://doi.org/10.1016/j.jecp.2011.12.005

Starkey, G. S., & McCandliss, B. D. (2014). The emergence of “groupitizing” in children’s numerical cognition. *Journal of Experimental Child Psychology*, *126*, 120–137. https://doi.org/10.1016/j.jecp.2014.03.006

Vogel, S. E., Koren, N., Falb, S., Haselwander, M., Spradley, A., Schadenbauer, P., Tanzmeister, S., & Grabner, R. H. (2019). Automatic and intentional processing of numerical order and its relationship to arithmetic performance. *Acta Psychologica*, *193*, 30–41. https://doi.org/10.1016/j.actpsy.2018.12.001

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, *138*(6), 1218–1252. https://doi.org/10.1037/a0029334

Wertheimer, M. (1922). Untersuchungen zur Lehre von der Gestalt. *Gestalt Theory*, *39*(1), 79–89. https://doi.org/10.1515/gth-2017-0007

Zar, J. H. (2010). *Biostatistical analysis* (5th edn). Pearson Prentice Hall.

Supplementary materials

1. ***Overview of studies examining perceptual grouping effects in arithmetic***

Table S1. Comparison of research on perceptual biases in processing arithmetic.

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
Main conclusions	Changes in spacing bias how people parse and solve expressions. More narrowly spaced chunks are treated as “go first,” often overriding formal order-of-operations. People give *larger* answers to more widely spaced problems (“longer looks larger”). Visual grouping and attention act like hidden gears inside symbolic reasoning.	With expressions like: a − b + c × d, using closer spacing of the last three terms and/or shifting the first term with the minus left makes people “detach” the minus (they solve it as: a − (b+c×d)). Shifting the minus to the right reduces those detachment errors. *Story problem generation task* (participants were asked to create a story corresponding to the given equations) and *Story problem solving task* (participants were asked to create an equation corresponding to the given story and to solve the story problem). In story problem generation task spatial tweaks also changed how people conceptually interpreted equations, not just how they computed.	The spacing effect survives when expressions are briefly flashed and masked (so participants rely on memory) and even when they verbalize expressions. Suggests a durable visuospatial representation of expressions; spacing can even bias symbol identity errors under tougher visual conditions.	Using >1.3M responses from Dutch primary students, two non-formal mechanisms drive order choice: perceptual grouping (closer terms grouped) and opportunistic selection (do the easy sub-problem). Crucially, these influences increase with grade level, challenging the “development = more abstract rule use” story. Spacing effects on order errors grew by roughly ~2 percentage points per grade.	In a large classroom RCT (grades 5–12, ASSISTments), incongruent spacing (additions tighter than multiplications) slowed mastery and increased errors. Congruent or neutral spacing led to faster mastery. Replicates lab effects in the wild; spacing meaningfully impacts real homework performance. HLM models controlling for grade and prior performance showed: -Prior performance predicted mastery speed (better prior → faster mastery). -Grade level was not a significant predictor. -The condition effect held: Congruent and neutral spacing led to faster mastery than incongruent spacing

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
		In contrast to the story-problem generation task, participants rarely made detachment errors when solving the given story problems, as the story context appeared to support a mathematically correct conceptual structure.			
Number of participants	Experiment 3: N=38 Indiana University undergraduate	N=91 undergraduates	Experiment 3: N= 21 undergraduate college students	50,000 children representing approximately 10% of all primary schools in the Netherlands ///collected from 65,856 unique Math Garden users over a period of 23 months	2,152 students (grades 5–12) doing online homework
Experiment environment (online/ onsite)	lab	lab	lab	Online – ed platform Math Garden website	Online – ed platform: the ASSISTments platform
Number of trials	200 / 45 minutes	18 problems/ 30 minutes	192	No fixed number of trials 1,5 mln trials/65856 users	No fixed number of trials 3 neutrally spaced and then randomly assigned: neutral, congruent, incongruent, mixed, spacing until they reach mastery (3 consecutive correct answers) on average 6 problems
Response format	Production paradigm	Production paradigm	Production paradigm	Production paradigm	Production paradigm
Age/ Education level	Undergraduate students	Undergraduate students	Undergraduate students	Grades 4-8, equivalent to Grades 2 to 6 (approximately aged 8-12) in the USA	equivalent to U.S. grade levels 2-6

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
Performance threshold / inclusion criteria	X	X	≥75% correct answers	No clear threshold or performance cutoff, but: - Only children who had already demonstrated competence with the basic arithmetic operations (+, –, ×, ÷) in earlier Math Garden games could play the *Arithmetic Sequence* game -The adaptive algorithm ensured that each child mostly saw problems at ~75% expected accuracy	Students had to complete more than three problems in the Skill Builder set. They also had to reach mastery, defined as three consecutive correct answers Students who quit early or completed within the first three problems (before seeing experimental spacing), were excluded. 2152 students out of 6,238 analysed - 4,053 students excluded (they stopped before finishing), 33 excluded due to unknown mastery status)
Operation type	Experiment 3: + , ×	+, -, x two types of equations: addition-first equations (a + b − c × d = __) and subtraction first equations (a − b + c × d = __) + story problems – generate a story problem to correspond with a given subtraction-first equation + given story problem and asked to write the corresponding equation and to solve the story problem (e.g., “Jake has a total of 45 dollars right now. He plans to pay 12 dollars for lunch today. Jake will also babysit	++, ××, +×, or ×+	+, x, -, / 308 target problems divided into 4 problem families (24, 54, 60 and 60 problems). Family 1 (Operator precedence, spacing manipulation) Problems with + and ×, testing whether narrow spacing biases order of operations. 60 problems total Family 2 (Left-to-right evaluation, spacing manipulation)	+, ×, mixed with –

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
		for two hours this evening and earn 7 dollars an hour. How much money will he have by tomorrow morning")		Problems with two subtractions or two divisions, testing whether spacing biases evaluation order. 60 problems total Family 3 (Operator precedence, ease of calculation manipulation) Problems with subtraction and multiplication, testing whether "easy" subtractions (e.g., sharing units digit) attract priority. 24 problems total Family 4 (Left-to-right evaluation, ease of calculation manipulation) Problems with two subtractions or two divisions, testing whether easy operations attract priority. 54 problems in total	
Measures	Exp 3: Response time -operation order -problem size - congruency Error rates -overall -precedence -operation/operator -operand (mistake in number)	% participants who produced detachment errors (percentage of participants in each condition who used each procedure at least once for each problem type, scored from their written work)	median response times, overall error rates, operator precedence error rates, and symbol identity error rates. (RT: the elapsed time between the presentation of the expression and the participant's next button press)	Foil error rate (Proportion of responses where children applied the *incorrect order of operations* (e.g., adding before multiplying). Calculated as number of foil errors ÷ total responses for each problem) Difficulty rating	the number of problems required to reach mastery (how many problems it took before three consecutive correct answers) Prior Mathematics Performance -ASSISTments Grade Level

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
		correct, detachment errors, left-to-right errors, and general errors analysis of written work of participants		Derived from the *Math Garden* adaptive algorithm Error rate (incorrect/ total responses).) Proportion of foil errors (Foil errors ÷ total errors)	
Conditions	2 (operation order: times – plus or plus – times) × 2 (spacing: narrow – wide versus wide – narrow) × 5 (problem size)	2 (operand spacing: evenly spaced or closely spaced) × 3 (first operator position: left shift, no shift, or right shift) between subjects design, yielding a total of six conditions Problem order was not manipulated (- first 10 eQ, THEN 8eq with +first)	Consistent vs inconsistent spacing with operator precedence EXP1: consistent vs. inconsistent spacing EXP2: mental arithmetic with masking (as EXP1, stimuli masked after 500ms) – forcing mental calculation EXP3 – immediate verbalization of the expression before solving	Family 1 – Operator precedence × Spacing Operations: Addition (+) and (×). Manipulation: Narrow spacing around plus vs. narrow spacing around times. Foil error = performing addition before multiplication. Family 2 – Left-to-right × Spacing Operations: Two subtractions or two divisions. Manipulation: Narrow spacing around the first operator vs. narrow spacing around the second operator. Foil error = performing the second operation first (violating left-to-right rule). Family 3 – Operator precedence × Ease of calculation	Neutral Congruent Incongruent Mixed Operation order

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
				Operations: Subtraction and Multiplication. Manipulation: Subtraction easy (shared units digit) vs. neutral subtraction. Foil error = performing subtraction first instead of multiplication. Family 4 – Left-to-right × Ease of calculation Operations: Two subtractions or two divisions. Manipulation: First operation easy vs. second operation easy. Foil error = executing the second operation first.	
Spacing	40 mm between operands for narrow problems (including the operation sign), 100 mm for widely spaced problems	one or two spaces between operands and operators for problems written in 12-point font. (narrow spacing was created by removing one space from the default two-space format)	Operand–operator spacing was measured from the center of the operator to the nearest number: wide spacing was ~0.78 deg and narrow spacing was ~0.45 deg of visual angle.	Specific values not reported	Specific values not reported (no extra spaces in neutral conditions)
Statistical methods	ANOVA	Logistic regression	Wilcoxon signed-rank test → comparison of error rates (inconsistent vs. consistent spacing). t-tests → comparison of median RTs across conditions.	Item-level analysis (not participant-level analysis), since each student saw only a subset of the 308 problems. Repeated-measures ANOVA tested whether manipulations (spacing or	A one-way ANOVA with condition (neutral, congruent, incongruent, and mixed) as the independent variable and mastery speed as the dependent measure. Post hoc tests with Bonferroni correction to examine where there were

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
			Effect sizes: *r* and Cohen's *d*.	ease of calculation) affected difficulty and foil error rate. Grade level was added as a predictor to test for developmental trends	significant differences in average mastery speed between conditions Hierarchical Linear Models (HLM, multilevel analysis) Used because ~10% of variance in mastery speed was attributable to classroom-level nesting (intraclass correlation ≈ 0.10). Level-1 (student level) predictors: Grade level (treated as continuous), Prior performance (proportion correct on previous ASSISTments work), Condition (dummy-coded, incongruent as reference group). Level-2 (teacher/class level) accounted for classroom nesting. The models tested: Model 2: Effects of grade level and prior performance. Model 3: Added spacing condition (neutral, congruent, mixed vs. incongruent). Model 4: Added interaction between prior performance × condition.

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
					Model 5: Added interaction between grade level × condition
Problem size	Defined by larger operand, analysed for 5 options: 2,3,6,7,8	X	X	In family 3 and 4, simple vs neutral (3: the size of subtraction, 4: easy first of second operation)	X
Errors	operation errors (+,× substituted), operand errors (wrong number), and precedence errors.	detachment errors (for „-„ only on first position) left-to-right errors (those errors did not appear in closely spaced conditions) other errors	Operator precedence errors Symbol identity errors (treating + as × nd vice versa) Other errors Overall error rates	Foil errors: the problems were designed so that executing operations in the incorrect order would always produce an incorrect response called 'foil error'.) "Difficulty ratings and foil error rates were both significantly higher when narrow spacing was used around operators that should have been executed second, rather than first", Resutls section, p. 22)	X
Overall accuracy	Reported for experiment 1 (95%±0,7%), not reported for experiment 3	Not reported explicitly Error analysis reported	Overall accuracy Exp. 1 (visible): Consistent-spacing trials ≈ 97% correct (so ≈3% overall errors). Exp. 2 (masked, 500 ms + mask): Consistent-spacing trials ≈ 93% correct (both masks), i.e., ≈7% overall errors. Exp. 3 (verbalization): No single % correct is stated; errors are analyzed by type	Not reported explicitly Error rates reported	X

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
Artithmetical fluency	X	X	X	Adaptive algorithm / rating to adjust problem difficulty to the level of a participant	Prior performance in ASSISTments as a covariate
Addition to Landy & Goldstone (2010)	X	examining written work examining minus sign	Masking and verbalisation paradigms	Children in different grades More naturalistic context 4 families of problems (subtraction, division)	Children in different grades Minus sign More naturalistic classroom context
Theoretical framework	"People incorporate spacing into operation ordering even when it is only occasionally present and generally unreliable." (p. 1965) Spacing is used as a cue to operation type – computation may be affected by the layout of problem presentation. "In contrast to Models of single-problem arithmetic (Campbell, 1994; Dehaene & Cohen, 1995; McCloskey & Lindemann, 1992), as well as models of multiterm computation (Anderson, 2005; Koedinger & MacLaren, 1997)" authors assumed that "single problem computation is independent of abstract problem structure." (p.1965)	Frames their findings as evidence that spatial layout biases cognitive processing through low-level perceptual mechanisms rather than purely symbolic reasoning	Gestalt principles of perceptual organization (proximity) – spacing acts as a grouping cue. Perceptual grouping and attention in arithmetic cognition – building directly on Landy & Goldstone (2007a, 2010), who argued that operator precedence is enforced partly through low-level perceptual processes like unit formation and attentional biases Hybrid representation perspective – they contrasted verbal working memory accounts of mental arithmetic (phonological loop, articulatory suppression studies) with their finding that even when expressions are masked or verbalized, visuospatial characteristics persist in mental representations of arithmetic	Perceptual grouping: a tendency to group spatially close symbols. It affects how people process information. "Evaluation of complex expressions begins with identification of simpler sub-expressions, but perceptual constraints rather than formal rules determine which symbols are grouped together to form sub-expressions." (p. 5) "Opportunistic selection refers to prioritizing for evaluation sub-expressions which are relatively easy to evaluate." (p. 6) "Subtraction and division operations were more likely to be evaluated prematurely if doing so made them simpler to evaluate (Families 3–4)." (p. 22)	Gestalt principles of grouping (proximity): Terms spaced closely tend to be grouped, which biases how students parse and solve expressions Perceptual learning theory (Gibson, 1969): They argue that perception adapts to highlight relevant features of notation, supporting higher-level cognition Perceptual influence on mathematical reasoning: Authors cite prior work (Landy & Goldstone, 2007; Jiang et al., 2014; Rivera & Garrigan, 2016) showing that spacing can bias students to apply incorrect order of operations. Thus, mathematical reasoning is at least partly perceptually driven, not purely rule-based Developmental/educational perspective:

Article	(Landy & Goldstone, 2010)	(Jiang et al., 2014)	(Rivera & Garrigan, 2016)	(Braithwaite et al., 2016)	(Harrison et al., 2020)
	"When perceptual processing of an item is manipulated, people are sensitive to the resulting psychological consequences on their performance and end up incorrectly attributing the basis for their performance consequences." (p.1966)				Authors argue that they extended this framework to classroom settings, suggesting that perceptual grouping effects are robust enough to appear in real homework contexts, not only in labs.

2. The example stimuli and complete stimuli set

Figure S1. A screenshot illustrating the example stimuli. a screenshot illustrating the example stimuli

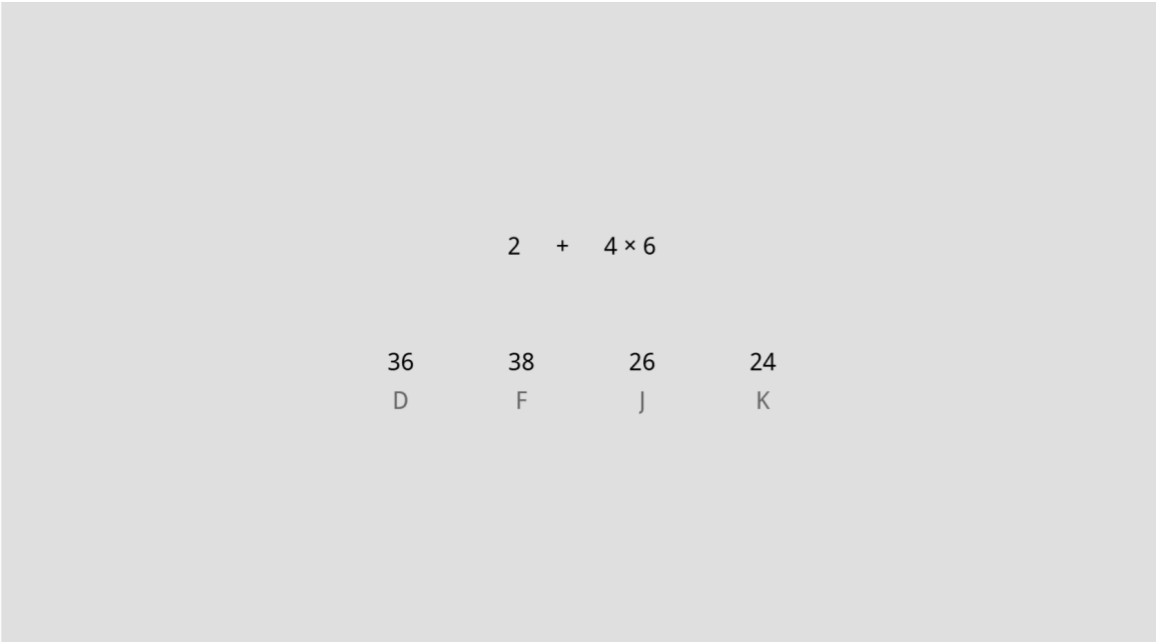


Table S2. Complete stimulus set.

All participants completed the same set of 200 arithmetic expressions. The design comprised four within-subject conditions (2 Congruency × 2 Order), with 50 trials per condition, yielding a total of 200 trials per participant. The two levels of Order were addition-first (a + b × c) and multiplication-first (a × b + c), and the two levels of Congruency were congruent spacing (operands of the multiplication operation presented closer together) and incongruent spacing (operands of the addition operation presented closer together).

Each row of the table corresponds to a single trial and specifies the structure of the arithmetic expression and response options. Columns are as follows: ***Digit 1, Oper 1, Digit 2, Oper 2, Digit 3*** *define the presented expression;* ***Correct response*** *indicates the correct solution;* ***Congruency*** *codes congruent and incongruent trials;* ***Ans 1–Ans 4*** *list the four response alternative presented to participant; Order* ***error resp*** *codes the order error response and* ***Distract resp*** *denotes how the predefined distractor alternatives were created (smaller[-] or larger [+] by two than correct response). Note that spacing between operands isn't presented explicitly in the table, but is defined by congruent and incongruent column.*

Digit 1	Oper 1	Digit 2	Oper 2	Digit 3	Correct response	Congruency	Ans 1 d	Ans 2 f	Ans 3 j	Ans 4 k	Order error resp	Distract response
2	+	3	×	2	d	con	8	6	12	10	k	--
2	×	3	+	2	f	con	6	8	10	12	j	--
2	+	4	×	2	j	con	14	12	10	8	f	--
2	×	4	+	2	k	con	12	8	14	10	d	--
2	+	3	×	3	d	con	11	13	15	17	j	++
2	×	3	+	3	f	con	12	9	11	14	d	++
2	+	4	×	3	j	con	20	16	14	18	k	++
2	×	4	+	3	k	con	16	14	13	11	f	++
2	+	3	×	6	d	con	20	30	22	28	f	+-
2	×	3	+	6	f	con	14	12	16	18	k	+-
2	+	4	×	6	j	con	36	38	26	24	d	+-
2	×	4	+	6	k	con	22	12	20	14	j	+-

2	+	3	×	8	d	con	26	24	42	40	k	-+
2	×	3	+	8	f	con	12	14	22	24	j	-+
2	+	4	×	8	j	con	46	48	34	38	f	-+
2	×	4	+	8	k	con	24	22	18	16	d	-+
2	+	3	×	9	d	con	29	27	45	43	j	--
2	×	3	+	9	f	con	24	15	22	13	d	--
2	+	4	×	9	j	con	36	52	38	54	k	--
2	×	4	+	9	k	con	24	26	15	17	f	--
3	+	3	×	2	d	con	9	12	11	14	f	++
3	×	3	+	2	f	con	13	11	17	15	k	++
3	+	4	×	2	j	con	14	16	11	13	d	++
3	×	4	+	2	k	con	20	16	18	14	j	++
3	+	3	×	3	d	con	12	14	16	18	k	+-
3	×	3	+	3	f	con	14	12	18	16	j	+-
3	+	4	×	3	j	con	23	21	15	11	f	+-
3	×	4	+	3	k	con	21	23	11	15	d	+-
3	+	3	×	6	d	con	21	19	36	38	j	-+
3	×	3	+	6	f	con	27	15	25	17	d	-+
3	+	4	×	6	j	con	25	44	27	42	k	-+
3	×	4	+	6	k	con	28	30	20	18	f	-+
3	+	3	×	8	d	con	27	48	25	44	f	--
3	×	3	+	8	f	con	15	17	31	33	k	--
3	+	4	×	8	j	con	56	54	35	33	d	--
3	×	4	+	8	k	con	34	18	36	20	j	--
3	+	3	×	9	d	con	30	32	56	54	k	++
3	×	3	+	9	f	con	20	18	36	38	j	++
3	+	4	×	9	j	con	65	63	39	41	f	++
3	×	4	+	9	k	con	39	41	23	21	d	++
6	+	3	×	2	d	con	12	14	18	16	j	+-

6	×	3	+	2	f	con	30	20	32	18	d	+-
6	+	4	×	2	j	con	16	18	14	20	k	+-
6	×	4	+	2	k	con	38	36	24	26	f	+-
6	+	3	×	3	d	con	15	27	13	29	f	-+
6	×	3	+	3	f	con	19	21	38	36	k	-+
6	+	4	×	3	j	con	30	28	18	20	d	-+
6	×	4	+	3	k	con	40	29	42	27	j	-+
6	+	3	×	6	d	con	24	22	52	54	k	--
6	×	3	+	6	f	con	22	24	54	52	j	--
6	+	4	×	6	j	con	58	60	30	28	f	--
6	×	4	+	6	k	con	60	58	28	30	d	--
6	+	3	×	8	d	con	30	32	72	74	j	++
6	×	3	+	8	f	con	66	26	68	28	d	++
6	+	4	×	8	j	con	40	82	38	80	k	++
6	×	4	+	8	k	con	74	72	34	32	f	++
6	+	3	×	9	d	con	33	81	35	79	f	+-
6	×	3	+	9	f	con	29	27	70	72	k	+-
6	+	4	×	9	j	con	90	92	42	40	d	+-
6	×	4	+	9	k	con	80	31	78	33	j	+-
8	+	3	×	2	d	con	14	12	26	22	k	-+
8	×	3	+	2	f	con	24	26	40	42	j	-+
8	+	4	×	2	j	con	22	24	16	18	f	-+
8	×	4	+	2	k	con	48	46	36	34	d	-+
8	+	3	×	3	d	con	17	15	33	31	j	--
8	×	3	+	3	f	con	48	27	46	25	d	--
8	+	4	×	3	j	con	18	34	20	36	k	--
8	×	4	+	3	k	con	54	56	33	35	f	--
8	+	3	×	6	d	con	26	66	28	68	f	++
8	×	3	+	6	f	con	32	30	74	72	k	++

8	+	4	×	6	j	con	72	74	32	34	d	++
8	×	4	+	6	k	con	82	40	80	38	j	++
8	+	3	×	8	d	con	32	34	86	88	k	+-
8	×	3	+	8	f	con	34	32	88	86	j	+-
8	+	4	×	8	j	con	98	96	40	38	f	+-
8	×	4	+	8	k	con	96	98	38	40	d	+-
8	+	3	×	9	d	con	35	33	99	101	j	-+
8	×	3	+	9	f	con	96	33	94	35	d	-+
8	+	4	×	9	j	con	42	110	44	108	k	-+
8	×	4	+	9	k	con	102	104	43	41	f	-+
9	+	3	×	2	d	con	15	24	13	22	f	--
9	×	3	+	2	f	con	27	29	43	45	k	--
9	+	4	×	2	j	con	26	24	17	15	d	--
9	×	4	+	2	k	con	52	36	54	38	j	--
9	+	3	×	3	d	con	18	20	38	36	k	++
9	×	3	+	3	f	con	32	30	54	56	j	++
9	+	4	×	3	j	con	41	39	21	23	f	++
9	×	4	+	3	k	con	63	65	41	39	d	++
9	+	3	×	6	d	con	27	29	72	70	j	+-
9	×	3	+	6	f	con	81	33	83	31	d	+-
9	+	4	×	6	j	con	35	76	33	78	k	+-
9	×	4	+	6	k	con	92	90	40	42	f	+-
9	+	3	×	8	d	con	33	96	31	98	f	-+
9	×	3	+	8	f	con	33	35	101	99	k	-+
9	+	4	×	8	j	con	104	102	41	43	d	-+
9	×	4	+	8	k	con	106	48	108	44	j	-+
9	+	3	×	9	d	con	36	34	106	108	k	--
9	×	3	+	9	f	con	34	36	108	106	j	--
9	+	4	×	9	j	con	115	117	45	43	f	--

9	×	4	+	9	k	con	117	115	43	45	d	--
2	+	3	×	2	d	incon	8	10	12	10	k	++
2	×	3	+	2	f	incon	10	8	10	12	j	++
2	+	4	×	2	j	incon	14	12	10	8	f	+-
2	×	4	+	2	k	incon	12	14	8	10	d	+-
2	+	3	×	3	d	incon	11	13	15	17	j	++
2	×	3	+	3	f	incon	12	9	14	11	d	++
2	+	4	×	3	j	incon	12	20	14	18	k	-+
2	×	4	+	3	k	incon	16	14	9	11	f	+-
2	+	3	×	6	d	incon	20	30	22	28	f	+-
2	×	3	+	6	f	incon	10	12	20	18	k	-+
2	+	4	×	6	j	incon	36	34	26	28	d	-+
2	×	4	+	6	k	incon	18	16	20	14	j	-+
2	+	3	×	8	d	incon	26	24	38	40	k	--
2	×	3	+	8	f	incon	12	14	22	20	j	--
2	+	4	×	8	j	incon	46	48	34	32	f	--
2	×	4	+	8	k	incon	24	22	14	16	d	--
2	+	3	×	9	d	incon	29	31	45	47	j	++
2	×	3	+	9	f	incon	24	15	26	17	d	++
2	+	4	×	9	j	incon	40	56	38	54	k	++
2	×	4	+	9	k	incon	28	26	19	17	f	++
3	+	3	×	2	d	incon	9	12	11	10	f	+-
3	×	3	+	2	f	incon	9	11	17	15	k	-+
3	+	4	×	2	j	incon	14	16	11	9	d	+-
3	×	4	+	2	k	incon	20	12	18	14	j	+-
3	+	3	×	3	d	incon	12	14	16	18	k	+-
3	×	3	+	3	f	incon	10	12	18	20	j	-+
3	+	4	×	3	j	incon	19	21	15	17	f	-+
3	×	4	+	3	k	incon	21	19	17	15	d	-+

3	+	3	×	6	d	incon	21	19	36	34	j	--
3	×	3	+	6	f	incon	27	15	25	13	d	--
3	+	4	×	6	j	incon	25	40	27	42	k	--
3	×	4	+	6	k	incon	28	30	16	18	f	--
3	+	3	×	8	d	incon	27	48	29	50	f	++
3	×	3	+	8	f	incon	19	17	35	33	k	++
3	+	4	×	8	j	incon	56	58	35	37	d	++
3	×	4	+	8	k	incon	38	22	36	20	j	++
3	+	3	×	9	d	incon	30	28	52	54	k	+-
3	×	3	+	9	f	incon	20	18	36	34	j	+-
3	+	4	×	9	j	incon	65	63	39	37	f	+-
3	×	4	+	9	k	incon	39	41	19	21	d	+-
6	+	3	×	2	d	incon	12	10	18	20	j	-+
6	×	3	+	2	f	incon	30	20	28	22	d	-+
6	+	4	×	2	j	incon	12	22	14	20	k	-+
6	×	4	+	2	k	incon	34	36	28	26	f	-+
6	+	3	×	3	d	incon	15	27	13	25	f	--
6	×	3	+	3	f	incon	19	21	34	36	k	--
6	+	4	×	3	j	incon	30	28	18	16	d	--
6	×	4	+	3	k	incon	40	25	42	27	j	--
6	+	3	×	6	d	incon	24	26	56	54	k	++
6	×	3	+	6	f	incon	26	24	54	56	j	++
6	+	4	×	6	j	incon	62	60	30	32	f	++
6	×	4	+	6	k	incon	60	62	32	30	d	++
6	+	3	×	8	d	incon	30	32	72	70	j	+-
6	×	3	+	8	f	incon	66	26	68	24	d	+-
6	+	4	×	8	j	incon	40	78	38	80	k	+-
6	×	4	+	8	k	incon	74	72	30	32	f	+-
6	+	3	×	9	d	incon	33	81	31	83	f	-+

6	×	3	+	9	f	incon	25	27	74	72	k	-+
6	+	4	×	9	j	incon	90	88	42	44	d	-+
6	×	4	+	9	k	incon	76	35	78	33	j	-+
8	+	3	×	2	d	incon	14	12	20	22	k	--
8	×	3	+	2	f	incon	24	26	40	38	j	--
8	+	4	×	2	j	incon	22	24	16	14	f	--
8	×	4	+	2	k	incon	48	46	32	34	d	--
8	+	3	×	3	d	incon	17	19	33	35	j	++
8	×	3	+	3	f	incon	48	27	50	29	d	++
8	+	4	×	3	j	incon	22	38	20	36	k	++
8	×	4	+	3	k	incon	58	56	39	35	f	++
8	+	3	×	6	d	incon	26	66	28	64	f	+-
8	×	3	+	6	f	incon	32	30	70	72	k	+-
8	+	4	×	6	j	incon	72	74	32	30	d	+-
8	×	4	+	6	k	incon	82	36	80	38	j	+-
8	+	3	×	8	d	incon	32	30	90	88	k	-+
8	×	3	+	8	f	incon	30	32	88	90	j	-+
8	+	4	×	8	j	incon	94	96	40	42	f	-+
8	×	4	+	8	k	incon	96	94	42	40	d	-+
8	+	3	×	9	d	incon	35	33	99	97	j	--
8	×	3	+	9	f	incon	96	33	94	31	d	--
8	+	4	×	9	j	incon	42	106	44	108	k	--
8	×	4	+	9	k	incon	102	104	39	41	f	--
9	+	3	×	2	d	incon	15	24	17	26	f	++
9	×	3	+	2	f	incon	31	29	47	45	k	++
9	+	4	×	2	j	incon	26	28	17	19	d	++
9	×	4	+	2	k	incon	56	40	54	38	j	++
9	+	3	×	3	d	incon	18	20	34	36	k	+-
9	×	3	+	3	f	incon	32	30	54	52	j	+-

9	+	4	×	3	j	incon	41	39	21	19	f	+-
9	×	4	+	3	k	incon	63	65	37	39	d	+-
9	+	3	×	6	d	incon	27	25	72	74	j	-+
9	×	3	+	6	f	incon	81	33	79	35	d	-+
9	+	4	×	6	j	incon	31	80	33	78	k	-+
9	×	4	+	6	k	incon	88	90	44	42	f	-+
9	+	3	×	8	d	incon	33	96	31	94	f	--
9	×	3	+	8	f	incon	33	35	97	99	k	--
9	+	4	×	8	j	incon	104	106	41	43	d	++
9	×	4	+	8	k	incon	110	46	108	44	j	++
9	+	3	×	9	d	incon	36	38	106	108	k	+-
9	×	3	+	9	f	incon	38	36	108	106	j	+-
9	+	4	×	9	j	incon	115	117	45	47	f	-+
9	×	4	+	9	k	incon	117	115	47	45	d	-+

3. *Untransformed accuracy results*

A 2 × 2 repeated measures ANOVA (analysis of variance) was conducted to examine the effects of Congruency (Congruent vs. Incongruent) and Order (Addition-First vs. Multiplication-First) on untransformed accuracy scores. There was a significant main effect of Congruency, $F(1, 200) = 23.96$, $p < .001$, $\eta^2_p = 0.11$, indicating that participants were more accurate in congruent spacing ($M = 0.92$, $SE = 0.01$) than in incongruent spacing ($M = 0.83$, $SE = 0.02$). A significant main effect of Order was also found, $F(1, 200) = 4.04$, $p = .046$, $\eta^2_p = 0.02$, suggesting higher performance for the Multiplication-First ($M = 0.88$, $SE = 0.01$) over the Addition-First ($M = 0.87$, $SE = 0.01$). The interaction between Congruency and Order was not significant, $F(1, 200) = 0.66$, $p = .418$, $\eta^2_p < 0.01$.

Outliers

There were participants whose overall accuracy was above the chance level (> 25%) but who exhibited near-zero accuracy within individual conditions. Majority of these participants (21 out of 25) appeared to understand the precedence rule – demonstrating above-chance performance in the congruent condition regardless of operation order – yet performed below chance in the incongruent condition. Therefore, we retained them in the analysis. These participants' accuracy was 85.43% (SE = 3.25%) in the congruent condition and 4.14% (SE = 1.21%) in the incongruent condition.

However, when in an additional analysis we excluded participants who scored below 25% accuracy in any condition, 176 participants remain in the dataset, no effects on untransformed accuracy remain significant. The main effect of Congruency was not significant, $F(1, 175) = 0.90$, $p = .34$, $\eta^2_p < .01$. The main effect of Order was not significant, $F(1, 175) < 0.01$, $p = .99$, $\eta^2_p < .01$.

4. ***Additional correlations: Relations between M4S performance and error types in the Precedence and Proximity Task***

M4S total score was negatively correlated with overall operation-order error rate, $r(199) = -.16$, $p = .024$, 95% CI [−.21, −.02], as well as with total error rate, $r(199) = -.22$, $p = .002$, 95% CI [−.35, −.08]. These results were confirmed using Spearman correlations (operation-order errors: $\rho = -.19$, $p = .007$; total errors: $\rho = -.21$, $p = .002$).

Importantly, the magnitude of the correlation between M4S and operation-order errors was numerically smaller than the correlation with overall accuracy reported in the manuscript ($r = .21$, $p = 0.002$) even though CI largely overlap, and substantially weaker than the correlation with RTs ($r = -.64$, $p < .001$). This indicates that while arithmetic fluency (as measured by M4S) is associated with fewer errors, its relationship with operation-order errors specifically is relatively small. This may be partly due to relatively low variance in the accuracy in the task.

Additionally, we observed a positive correlation between operation-order error rates in the Precedence and Proximity Task and calculation error rates in M4S ($r = .31$, $p < .001$), suggesting partial overlap between general calculation difficulty and susceptibility to operation-order violations.

Overall, these results suggest that M4S captures general arithmetic proficiency that relates to both speed and accuracy, whereas operation-order errors in the Precedence and Proximity Task are only weakly associated with this measure. This pattern is consistent with the interpretation that operation-order errors are not solely driven by general calculation ability, but may also reflect task-specific cognitive mechanisms (e.g., perceptual grouping effects). At the same time, these findings should be interpreted cautiously because error rates, especially operation-order errors, were generally low, which constrains the stability and interpretability of correlational estimates.

References

Braithwaite, D. W., Goldstone, R. L., Van Der Maas, H. L. J., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition*, *151*, 113. https://doi.org/10.1016/j.cognition.2016.03.024

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The Education of Perception. *Topics in Cognitive Science*, *2*(2), 265–284. https://doi.org/10.1111/j.1756-8765.2009.01055.x

Harrison, A., Smith, H., Hulse, T., & Ottmar, E. R. (2020). Spacing Out! Manipulating Spatial Features in Mathematical Expressions Affects Performance. *Journal of Numerical Cognition*, *6*(2), 186–203. https://doi.org/10.5964/jnc.v6i2.243

Jiang, M. J., Cooper, J. L., & Alibali, M. W. (2014). Spatial factors influence arithmetic performance: The case of the minus sign. *Quarterly Journal of Experimental Psychology (2006)*, *67*(8), 1626–1642. https://doi.org/10.1080/17470218.2014.898669

Landy, D., & Goldstone, R. L. (2010). Proximity and Precedence in Arithmetic. *Quarterly Journal of Experimental Psychology*, *63*(10), 1953–1968. https://doi.org/10.1080/17470211003787619

Rivera, J., & Garrigan, P. (2016). Persistent perceptual grouping effects in the evaluation of simple arithmetic expressions. *Memory & Cognition*, *44*(5), 750–761. https://doi.org/10.3758/s13421-016-0593-z