HHS Public Access

Author manuscript

J Exp Child Psychol. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

J Exp Child Psychol. 2022 July ; 219: 105399. doi:10.1016/j.jecp.2022.105399.

When multiplying is meaningful in memory: Electrophysiological signature of the problem size effect in children

Danielle S. Dickson[1,2], **Amandine E. Grenier**[1], **Bianca O. Obinyan**[3], **Nicole Y.Y. Wicha**[*]

Department of Neuroscience, Developmental and Regenerative Biology, The University of Texas at San Antonio, San Antonio, TX 78249, USA

Abstract

Children are less fluent at verifying the answers to larger single-digit arithmetic problems compared with smaller ones. This problem size effect may reflect the structure of memory for arithmetic facts. In the current study, typically developing third to fifth graders judged the correctness of single-digit multiplication problems, presented as a sequence of three digits, that were either small (e.g., 4 3 12 vs. 4 3 16) or large (e.g., 8 7 56 vs. 8 7 64). We measured the N400, an index of access to semantic memory, along with accuracy and response time. The N400 was modulated by problem size only for correct solutions, with larger amplitude for large problems than for small problems. This suggests that only solutions that exist in memory (i.e., correct solutions) reflect a modulation of semantic access likely based on the relative frequency of encountering small versus large problems. The absence of an N400 problem size effect for incorrect solutions suggests that the behavioral problem size effects were not due to differences in initial access to memory but instead were due to a later stage of cognitive processing that was reflected in a post-N400 main effect of problem size. A second post-N400 main effect of correctness at occipital electrodes resembles the beginning of an adult-like brain response observed in prior studies. In sum, event-related brain potentials revealed different cognitive processes for correct and incorrect solutions. These results allude to a gradual transition to an adult-like brain response, from verifying multiplication problems using semantic memory to doing so using more automatic categorization.

Keywords

N400; Multiplication verification; Problem size; Typically developing children; Semantic memory; ERP

[*]Corresponding author. nicole.wicha@utsa.edu (N.Y.Y. Wicha).

[2]Current address: Department of Psychology and Center for Language Science, The Pennsylvania State University, State College, PA 16801, USA.

[3]Current address: School of Medicine, The University of Texas Medical Branch at Galveston, Galveston, TX 77555, USA.

[1]These authors contributed equally to this work as co-first authors.

Introduction

General

Children's ability to perform elementary arithmetic provides a foundation for their later acquisition of higher-level mathematical skills in high school and college. As arithmetic fluency increases, reliance on relatively inefficient processes, such as counting and decomposition, declines in favor of memory-based retrieval (Ashcraft, 1982; Lemaire & Siegler, 1995; Siegler, 1988), and retrieval deficits have been associated with mathematical disability (Geary, Hoard, & Bailey, 2012; LeFevre, Daley, et al., 1996). Successful and efficient retrieval is dependent on the development and maintenance of accurate long-term memory representations for arithmetic facts such as the correct answers to single-digit multiplication problems. Characterizing the quality and retrieval processes of these memories, especially when they are first being established, is important both at a theoretical level, for resolving debates about cognitive arithmetic and its relationship (or not) to language processing (e.g., Butterworth, Reeve, Reynolds, & Lloyd, 2008; Frank, Everett, Fedorenko, & Gibson, 2008; Gelman & Gallistel, 2004), and at a practical level, for understanding potential vulnerabilities in the system.

Children who are first learning core arithmetic skills begin their education by operating on smaller numerosity and then progress to working with larger numerical values. As a result, effects of problem size, such as responding slower to 8×7 than to 4×3, have been observed. As arithmetic fluency increases, problem size effects gradually decrease but continue to persist into adulthood (for reviews, see Ashcraft & Guillaume, 2009; Campbell & Graham, 1985; De Brauwer, Verguts, & Fias, 2006; Dickson & Wicha, 2019; Koshmider & Ashcraft, 1991; Zbrodoff & Logan, 2005). Indeed, the problem size effect is so consistently observed across populations that one review referred to it as a phenomenon that "everyone finds" (Zbrodoff & Logan, 2005). The reduction in the problem size effect that occurs with increased arithmetic fluency is dependent on successful encoding of more difficult, typically larger problems into long-term memory (see Ashcraft & Guillaume, 2009, for additional perspectives). However, the source of this effect is still debated.

In adults, the explanation for the problem size effect has focused on three interrelated factors: frequency of occurrence, interference or "confusion" in memory, and strategy. First, in education, small problems are learned earlier and encountered more frequently (Ashcraft & Christy, 1995), leading to stronger memory representations and faster retrieval for smaller problems than for larger problems (Imbo & Vandierendonck, 2008; Zbrodoff & Logan, 2005). Second, it has been proposed that answer retrieval for larger problems may also involve more interference or "confusion" in memory (for a review, see Ashcraft & Guillaume, 2009). This interference may result from larger problems having a greater history of retrieval errors during learning, as compared with smaller problems, which are easier to remember and less prone to retrieval errors. Third, procedural strategies for solving arithmetic problems, such as repeated addition or transformation, are reportedly more frequent for larger problems than for smaller problems (cf. Kirk & Ashcraft, 2001; LeFevre, Sadesky, & Bisanz, 1996; Siegler, 1988). Each of these may have different implications for memory representations and processing of arithmetic facts.

Traditionally, the processing of arithmetic facts has been inferred from behavioral measures of response time and accuracy using simple arithmetic tasks in which participants produce or verify the solution to a problem (Campbell & Austin, 2002; for reviews, see Domahs & Delazer, 2005; Noël, Fias, & Brysbaert, 1997). The prevailing understanding is that large, less well-learned problems are more weakly represented in memory and therefore, when these problems are presented, the solutions are not as accessible. However, the mechanisms involved at different levels/stages of arithmetic processing are not well understood; that is, does sensitivity to size affect access to correct and incorrect solutions equally, or are the behavioral effects of problem size that are observed in verification paradigms instead driven by downstream processes such as response selection and execution? We begin to address these questions in the current study.

Arithmetic processing in children

Children who have just learned simple arithmetic problems make for a particularly compelling study population because the memories they form during learning construct the foundation on which more advanced mathematical skills are built. Indeed, behavioral findings from children underpin much of the theoretical impetus for the approaches used in studying cognitive arithmetic more broadly (Siegler & Braithwaite, 2017). Notably, arithmetic knowledge in children has not yet been altered by disuse or learning additional skills in higher mathematics courses. Thus, by studying how children respond to arithmetic problems, we can test the functionality of the emerging arithmetic memory system and compare that with what is later used during adulthood (Imbo & Vandierendonck, 2008).

Determining when and how problem size, or problem difficulty, affects processing in children requires an understanding of the cognitive mechanisms that unfold prior to “cognitive end-state” accuracy and response time measures. That is, by the time a child makes a response to a problem, a cascade of cognitive events has transpired. Despite the importance of studying this, research using time-sensitive neuroimaging measures is surprisingly limited in children (Peters & De Smedt, 2018). Event-related brain potentials (ERPs) have proven to be useful as a window onto these cognitive mechanisms in children. The results from initial studies suggest that children may exhibit different brain responses than adults when processing multiplication problems (Cerda, Grenier, & Wicha, 2019; Grenier, Dickson, Sparks, & Wicha, 2020; Wicha, Dickson, & Martinez-Lincoln, 2018). This distinction, however, was not always recognized (Prieto-Corona et al., 2010), in part because even the adult brain response had not yet been well characterized (Dickson et al., 2018; Dickson & Wicha, 2019; Jasinski & Coch, 2012; Prieto-Corona et al., 2010). We briefly introduce this adult literature to set the historical context for understanding the neurocognitive indices for simple arithmetic in children.

ERP componentry during arithmetic verification tasks

An ERP is a measure of continuous electrical brain activity time-locked and averaged across multiple presentations of a stimulus of interest such as the solution to a multiplication problem. The multidimensional nature of ERP waveforms allows for inferences about the underlying cognitive processes based on the timing (latency), polarity (negative-going vs. positive-going morphology), amplitude (microvolts), and distribution across the scalp. ERPs

can even reveal changes in cognitive mechanisms that are not measurable by less sensitive endpoint behavioral responses (e.g., McLaughlin, Osterhout, & Kim, 2004) or can reveal distinct cognitive mechanisms for similar behavioral outcomes across populations (Grenier et al., 2020). The most relevant ERP effect for the current study emerges as a difference in voltage amplitude on a negative-going waveform, called the N400, which is associated with attempted access to semantic memory (Federmeier, 2022; Kutas & Federmeier, 2011; Kutas & Hillyard, 1984). N400 amplitude is reduced (i.e., less negative) with increasing ease of semantic access, particularly through priming or other means of contextual support. In other words, more expected items tend to elicit progressively less negative N400 responses. Critically, the N400 is an automatic brain response to any meaningful or potentially meaningful stimulus and can appear even without conscious awareness of the event (Luck, Vogel, & Shapiro, 1996).

Grenier et al. (2020) demonstrated that presenting children with multiplication problems that are either correct or incorrect (e.g., $2 \times 4 = 8$ vs. $2 \times 4 = 9$) modulates the N400 similarly to a language task (e.g., reading sentences with nouns that make sense or not in the context) (Kutas & Hillyard, 1984). In contrast, adults elicit a different ERP at the solution, namely a positive-going response called a P300, with larger amplitude for the correct solutions compared with the incorrect solutions (see also Dickson et al., 2018; Dickson & Wicha, 2019; Jasinski & Coch, 2012). The P300 is typically associated with stimulus categorization and is larger for task-relevant targets compared with distractors (Polich, 1987, 2007, 2012; Sutton, Braren, Zubin, & John, 1965; Sutton, Ruchkin, Munson, Kietzman, & Hammer, 1982). In arithmetic verification tasks, participants must make a categorical decision on the correctness of the presented solution. The P300 response in adults reveals that they are able to process the solutions to arithmetic problems efficiently, detecting the correct solution as their target (eliciting a large P300) and detecting incorrect solutions as distractors (Dickson & Wicha, 2019).

ERP studies have also reported an arithmetic problem size effect in adults, which has been interpreted as a late positive component (Niedeggen, Rösler, & Jost, 1999; Núñez-Peña, 2008; Szucs & Csépe, 2005) or as a delayed P300 (Dickson & Wicha, 2019). Much less is known about the ERP correlates of the problem size effect in children, with no studies to date testing verification (Van Beek, Ghesquière, De Smedt, & Lagae, 2014, 2015). Given that children and adults engage different cognitive processes—semantic access (N400) versus target categorization (P300) of the solutions, respectively— when verifying the correctness of simple multiplication problems (Cerda et al., 2019; Dickson et al., 2018; Dickson & Wicha, 2019; Grenier et al., 2020; Jasinski & Coch, 2012; Prieto-Corona et al., 2010; Wicha et al., 2018), the prediction here was that the arithmetic problem size effect in children would modulate the N400 as an index of differential access to smaller and larger problems in semantic memory.

Multiplication verification and problem size in children

The goal of the current study was to characterize the neural correlates for the problem size effect in children and, in turn, the cognitive process affected during a verification task. To do this, we measured the effect of problem size on the arithmetic N400 while typically

developing children verified the correctness of four types of multiplication solutions: correct and incorrect solutions to small problems (e.g., 2 × 4 = 8 or 10) and correct and incorrect solutions to large problems (e.g., 8 × 6 = 48 or 54). This same design was used in a separate study with adults (Dickson & Wicha, 2019), allowing for comparison of the overall ERP and behavioral patterns between populations. By considering the possible effects of problem size on correct and incorrect solutions separately, different questions could be addressed regarding how children process multiplication more broadly.

First, it is helpful to consider the broader literature on the N400 to better understand what the arithmetic N400 reflects. Grenier et al. (2020) showed that children elicit an N400 to multiplication solutions during a verification task, with larger negative amplitude for incorrect solutions than for correct solutions. This arithmetic N400 resembled the N400 elicited in a word–picture verification task and has been interpreted in line with decades of language comprehension research (Federmeier, 2022; Kutas & Federmeier, 2011). In sentence processing studies, a word that is expected based on context elicits smaller N400 amplitude than an unexpected word in that same context (e.g., “Dogs love to chew on bones/socks” (Federmeier & Kutas, 2001; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Kutas & Federmeier, 2011). Considering that smaller problems are easier to retrieve from memory, the correct solutions to small problems should in turn elicit smaller N400 amplitude than those to large problems, similar to semantically predictable words in a sentence context. This N400 amplitude modulation would reflect more facilitated access to semantic memory for small problems than for large problems.

How might proposed explanations for the problem size effect—frequency of encountering a problem, interference or “confusion” in memory, and strategy use—affect the arithmetic N400? In the language literature, words that are more frequently encountered in daily life also elicit smaller N400 amplitude compared with words that are less frequent (Fischer-Baum, Dickson, & Federmeier, 2014; Rugg, 1990; Van Petten, 1993). Words that have more competitors in memory elicit larger N400 amplitude (Meade, Grainger, Midgley, Emmorey, & Holcomb, 2018; Megías & Macizo, 2016). Therefore, regardless of whether frequency or confusion accounts for the problem size effect, correct solutions should elicit smaller N400 amplitude for small problems than for large problems.

The role of strategy in explaining the problem size effect is less clear. Although children quickly adopt retrieval strategies (Siegler, 1988), procedural processes are reportedly more frequent for larger problems than for smaller problems even in adults (cf. LeFevre, Daley, et al., 1996; Siegler, 1988). There is no evidence that the N400 is modulated by strategy per se (Federmeier, 2022; Kutas & Federmeier, 2011). There is, however, some evidence that tasks that demand more superficial or automatic processing of words can lead to the absence of an N400 modulation in a language task (Fischer-Baum et al., 2014). This is perhaps in line with adults eliciting a P300, not an N400, when verifying multiplication problems, reflecting their more superficial processing of information (Dickson et al., 2018; Dickson & Wicha, 2019; Grenier et al., 2020). At the same time, N400 amplitude is modulated when meaning level processing is required for the task even when the stimulus is not consciously perceived (Luck et al., 1996). In turn, if an N400 is observed in children for both small and large

problems, it is unlikely that differences in strategy would be an adequate explanation on its own for the problem size effect observed in behavior.

With regard to incorrect solutions, it is helpful to consider what is known from studies of problem size with adults. Unlike the larger N400 observed in children (Grenier et al., 2020), multiple studies have now demonstrated that when adults verify the correctness of simple multiplication problems, correct solutions elicit an earlier and larger positive-going ERP, or P300, than incorrect solutions (Dickson et al., 2018; Dickson & Wicha, 2019; Grenier et al., 2020; Jasinski & Coch, 2012). The P300 indexes the categorization of the solution as correct or incorrect. Consistent with the broader P300 literature, adults are so efficient at recognizing the correct solution to simple multiplication problems that they do not need to process the problems for meaning, as children do, and instead "detect" the correct solutions as "targets" and the incorrect solutions as "distractors" (Polich, 2012). Problem size further modulates the P300 for both correct and incorrect solutions in a graded fashion, with large incorrect solutions eliciting disproportionately smaller P300 amplitude compared with small incorrect and correct solutions. This is again consistent with the broader literature, revealing larger P300 amplitude to items that are easier to categorize (Dickson & Wicha, 2019; Isreal, Chesney, Wickens, & Donchin, 1980; Polich, 2007), and indicates that both problem size and correctness affect the ease with which adults can categorize the solutions.

In contrast to adults, the large modulation of the arithmetic N400 in children reveals that they process the solutions for meaning, with incorrect solutions being incongruous with expectations based on the preceding operands and eliciting a larger N400 amplitude (Grenier et al., 2020). One question is whether the brain response to both correct and incorrect solutions will be modulated by problem size as with adults. This is theoretically important because it has been argued that arithmetic facts are stored in a memory network as related facts, and as such related incorrect solutions should be more difficult to reject (Campbell, 1987; Campbell & Graham, 1985; Niedeggen & Rösler, 1999; Stazyk, Ashcraft, & Hamann, 1982). It would follow that incorrect solutions that are table related to the correct solution (e.g., $2 \times 4 = 12$) should be more available in semantic memory for smaller well-known problems than for larger problems that have weaker representations in semantic memory. In this case, an effect of problem size should reflect on ERP components measuring memory access, such as the N400, where greater ease of access leads to smaller N400 amplitude. Thus, incorrect solutions in small predictable problems should elicit smaller amplitude N400s compared with larger problems, reflecting spread of activation across this proposed memory network. We discuss other possible explanations of the problem size effect on incorrect solutions as they relate to the findings.

Lastly, with respect to possible ERP modulations other than the N400, there have been reports of problem size modulating a late positive component during an arithmetic production task in children (Van Beek et al., 2014, 2015). However, this later ERP effect has not been observed in verification of arithmetic facts (Cerda et al., 2019; Grenier et al., 2020). The current study should determine whether children elicit post-N400 effects that are modulated by problem size.

Method

Participants

A total of 57 children (24 female) in elementary Grades 3 to 5 were included in this study. These children were a subset of an existing large-scale study ($N = 99$) (Grenier et al., 2020). Prescreening for inclusion in that study included math fluency (i.e., minimum of a third-grade level) and task accuracy (i.e., above chance). Their original design examined the effect of solution correctness and included most single-digit multiplication problems. This allowed us to do a subsequent analysis of the effect of problem size. Children from the original study were included here if their data had the minimum number of ERP trials, as discussed below. For completeness, we compared the excluded children with the 57 children based on performance and demographic data only (see "Additional analysis" section in Results).

Children were recruited from the local community through word of mouth and advertisement (e.g., free family magazines, social media groups). All were right-handed and had normal or corrected-to-normal vision and normal hearing. Participation was contingent on their ability to meet a low threshold for basic arithmetic ability (Wechsler Individual Achievement Test–Third Edition [WIAT-III], described below). Participants who had a history of neurological abnormalities or trauma, had a documented language or mathematical disability (including attention-deficit/hyperactivity disorder, dyscalculia, or dyslexia), or were taking psychoactive medications were excluded.

Participants' average age was 10 years (range = 8 years 1 month to 11 years 9 months), and their average grade level was 4.7 (range = 3.7–5.9, where 5.9 covers the months of May–August prior to entering sixth grade in the fall), with 19 children in each grade (third, fourth, and fifth). The sample was representative of the diverse socioeconomic status (SES) and language backgrounds of the local population. The sample included 34 monolingual English speakers (no or limited exposure to a second language), 10 dominant English speakers (heritage speakers of Spanish or children enrolled in Spanish–English dual language programs), and 13 proficient Spanish–English bilinguals. All children had at least a low average level of English proficiency to ensure that they were able to follow instructions in English (vocabulary size and oral comprehension from Woodcock–Johnson III Tests of Achievement; Woodcock, McGrew, & Mather, 2001).[4]

SES was measured on the four-factor Hollingshead scale (full range = 8–66), which for children considers the highest level of education and occupation of their primary caretaker (Hollingshead, 1975). Participants' scores ranged from 16.5 to 66 and fell on average in upper-middle SES, with an average score of 50.

Informed consent (parent or legal guardian) and child assent were received in accordance with the institutional review board of the University of Texas at San Antonio (UTSA).

[4]Post hoc analyses were conducted to ensure that monolinguals ($n = 34$) and bilinguals ($n = 23$) did not differ based on performance (accuracy: $p = .65$; response time: $p = .72$) or ERP effects of problem size ($p = .38$).

Offline behavioral cognitive assessments

Prior to the electroencephalogram (EEG) recording, we obtained multiple offline measures of cognitive ability, which are summarized in Table 2 in Results. The most critical of these was proficiency on the multiplication subtest of the WIAT, which is a paper-and-pencil timed (1-min) production task with single-digit problems of increasing difficulty (math fluency–multiplication) (Wechsler, 2009). To ensure minimum competence on the task, children needed to meet a third-grade level for multiplication (sample range = 9–40 of 40 problems).

Stimuli

Single-digit multiplication problems consisted of three sequential Arabic numerals: two operands followed by a proposed solution. Because only one type of operation was presented, the multiplication sign was not included in the stimuli to keep the experiment shorter. All paired combinations for operands 2 through 9 were used except tie problems that are known to elicit reduced size effects given their unique status in memory (e.g., 3 × 3) (Miller, Perlmutter, & Keating, 1984). The resulting 56 problems were operationalized, half as small and half as large, by the size of their correct solution (see Stazyk et al., 1982). Large problems had correct solutions equal to or greater than 27, and small problems had correct solutions equal to or less than 24 (25 and 26 were not possible solutions in this set). Finally, across the experiment, each problem was presented once with a correct solution (e.g., 2 5 10) and once with an incorrect solution (e.g., 2 5 16) for a total of 112 trials.

Incorrect products were always multiples of one of the preceding operands (i.e., table-related solutions) and were generated by adding or subtracting 1 or 2 from one of the preceding operands [e.g., 2 × 5 => (2 + 1) × 5]. Zero or multiples of 10 or 11 were never used as possible incorrect solutions. A challenge with creating related incorrect problems is that incorrect solutions for large problems are naturally a further numerical distance from their correct solution than small incorrect solutions are from theirs. For example, the incorrect solution 5 × 3 = 12 is a distance of 3 from the correct solution 15, but 8 × 7 = 63 is a distance of 7 from the correct solution 56. However, by consistently adding or subtracting 1 or 2 from one operand, we avoided including easily discardable incorrect solutions for small problems. This way, incorrect solutions scale with the problem size without introducing the factor of plausibility, which could interfere with multiplication verification on the judgment task (Núñez-Peña & Escera, 2007; see Discussion). Unrelated solutions were not included in the stimuli in order to keep the length of the study tolerable for children. Table-related solutions were selected because they are more difficult to reject than unrelated solutions and therefore are more likely to elicit modulations on ERP components (Niedeggen & Rösler, 1999).

Procedure

A session began with the offline standardized measures and questionnaires, followed by the multiplication task with simultaneous measurement of EEG and performance (accuracy and response times). EEG and performance metrics were treated as separate dependent measures. During the experimental task, participants viewed the multiplication problems one operand at a time and were instructed to verify whether the third number was the correct

product for the first two operands. A short practice set preceded the task. Participants were instructed in English to respond as quickly and accurately as possible as soon as they saw the solution (responses were not delayed; cf. Dickson et al., 2018). Solution correctness was indicated on a Logitech F310 Gamepad (Newark, CA, USA) by button responses with participants' right and left index fingers. In the full sample, response-hand mapping was counterbalanced across participants by gender. In this sample, 31 (of 57) responded that a solution was correct with their right hand.

Paradigm (Perception Research Systems, Lawrence, KS, USA) was used to present the experimental stimuli in the center of a 19-inch LCD monitor positioned 100 cm away from participants. The presentation rate was based on Nieddegen and Rösler (1999) and was as follows. A 1000-ms cue to blink if needed (a cartoon eye) was followed by a 1000-ms cue (a yellow coin with a central X to remind participants that only multiplication problems were presented) to indicate that the next trial was starting and to encourage children to keep their eyes focused on the center of the monitor. Arabic numeral stimuli (80-point font) were presented in white text on a black background. Each operand was presented for 450 ms, with a 250-ms interstimulus interval (ISI) between them. The second operand was followed by a 1000-ms ISI, and then the proposed solution was presented for 350 ms. Following solution offset, a blank screen remained for a minimum of 1000 ms (to ensure an adequate ERP collection window, i.e., 1 s of post-stimulus data) and a maximum of 5000 ms. Button-press responses within that time window would immediately advance to the next trial. Trials with responses beyond the 5-s cutoff were not recorded. See Fig. 1 for an example of presentation and timing.

Trials were distributed across eight blocks of 14 problems each. To make the experiment child friendly, participants were told a brief cover story about having a mission on a rocket ship through a math universe where their goal was to collect as many coins as quickly as possible. A coin would be earned for each problem that was answered correctly, and these coins could be used to open different levels on a treasure box. Between blocks, the display informed the children of the total number of coins earned in that block (i.e., trials answered correctly) as well as the number of blocks completed and remaining. Although rewards can change performance (and can modulate performance-related ERPs such as the P300) (e.g., Carrillo-de-la-Peña & Cadaveira, 2000; Kleih, Nijboer, Halder, & Kübler, 2010), this feedback was critical for maintaining children's interest in the task. Children did not receive feedback on their accuracy on each trial, nor did they receive negative feedback (e.g., percentage correct) to avoid demotivation. Children exchanged their virtual coins for a toy prize at the end of the experiment. All children received a reward regardless of their performance, and their performance was never compared with that of other children.

EEG recording

EEG data collection—During the EEG recording, the child sat alone in a sound-attenuating, electrically shielded chamber and was monitored with a closed-circuit camera whenever the chamber door was closed. Continuous EEG was recorded from 26 Ag–AgCl active electrodes fitted into a geodesically arranged electrode cap (Electro-Cap International, Eaton, OH, USA) (see Fig. 3 in Results for configuration) using a BioSemi amplifier

running ActiveView software (BioSemi ActiveTwo; BioSemi B.V., Amsterdam). Recordings were also made from six external electrodes: under each eye to monitor blinking, at the outer canthus of each eye to monitor horizontal eye movements, and on the mastoid processes for referencing of the data. Electrode offsets were kept below 25 mV. An analog fixed first-order anti-aliasing filter with a half-power point at 3.6 kHz was applied (see https://www.biosemi.com). The data were sampled at 256 Hz (2048 Hz with a decimation factor of 1/8) with a digital low-pass fifth-order sinc response with a −3-dB point at 51.2 Hz (1/5 of the sample rate). Online recordings for each electrode were made with respect to common mode sense active and driven right leg passive electrodes.

EEG processing—Raw BioSemi data were imported into Version 14.1.1 of the MATLAB-based toolbox EEGLAB (Delorme & Makeig, 2004), which was used for processing and analysis in conjunction with Version 6.1.4 of the ERPLAB toolbox (Lopez-Calderon & Luck, 2014). The raw data were referenced to the average of the left and right mastoid electrodes. A high-pass second-order Butterworth filter with a 0.1-Hz cutoff was then applied to all channels. Epochs of raw EEG data were extracted with a 100-ms prestimulus baseline (−100 to 900 ms of stimulus onset). Trials with inaccurate behavioral responses were excluded, as were epochs contaminated with artifacts; blinks (see below), horizontal eye movements, excessive muscle artifacts, and channel drift were identified by individually calibrated thresholds using algorithms in ERPLAB. Participant-level average ERPs were generated, and a low-pass second-order Butterworth filter with a 30-Hz cutoff was then applied to these data prior to analysis.

As a factor of the original paradigm (Grenier et al., 2020), there was a maximum of 28 trials in each condition here. This left a small margin for data loss to maintain an adequate signal-to-noise ratio. Participants were included in ERP analyses if they had a minimum of 15 trials in each of the four conditions (average = 21 trials, range = 15–28). From the Grenier et al. (2020) dataset, 26 participants met these criteria after removing trials with artifacts, and an additional 31 participants met the criteria after independent component analysis (ICA) for blink correction. ICA was conducted using the EEGLAB runica algorithm over all 26 head channels on high-pass filtered continuous EEG data. Blink components were identified and removed using standard component spectra, scalp topography, and time course (Jung et al., 2000). Trials with artifacts were then removed from this modified dataset.

The average number of trials per critical bin was not significantly different across trial types (small correct = 22; large correct = 21; small incorrect = 21; large incorrect = 20). Table 1 shows the trial count distribution based on condition. Because of the nature of the task, there are more children with fewer trials in the (harder) incorrect large type and more trials in the (easier) correct small type. However, there was no significant difference in the number of trials on average across conditions.

ERP data analysis—All statistical analyses were performed in R (R Development Core Team, 2016). ERPs were measured separately from the onset of the second operands (problem size) and the solutions (problem size and correctness), each relative to a 100-ms prestimulus baseline. Separate analyses of variance (ANO-VAs) (using the *ez* package; Lawrence, 2016) were conducted for the second operand and solution with 2 levels of

problem size or problem size and correctness, respectively, and 26 levels of electrode. Significant interactions with electrode were subjected to additional tests for distributional analysis, with a subset of 16 electrodes that could be divided by anteriority (prefrontal, frontal, central, or posterior), hemisphere (left or right), and laterality (lateral or medial). Tests involving more than 1 degree of freedom in the numerator (e.g., anteriority) were corrected for violations of sphericity, and corrected p values and corresponding ε correction factors are reported. Generalized eta squared $\left(\eta^2_G\right)$ (Bakeman, 2005) is reported for effects that reach significance. When possible, effect contrasts are reported to provide within-participant differences and standard errors as well as their effect size (d_z).

Behavioral analysis—Response times (RTs) from trials that elicited accurate responses, regardless of whether the trials had usable corresponding ERPs, were included in analyses after removing extreme values (<200 ms or > 5 s; 8 trials removed). These RTs were subjected to an ANOVA with factors of size (small or large) and correctness (correct or incorrect). Percentage accuracy per critical condition was similarly assessed.

Results

Behavior

Results from the standardized measures are reported in Table 2.

Response times

Main effects of correctness, $F(1, 56) = 58.50$, $p < .001$, $\eta^2_G = .04$, and size, $F(1, 56) = 96.73$, $p < .001$, $\eta^2_G = .13$, reached significance. The interaction between these factors did not, $F(1, 56) = 2.43$, $p = .12$. Responses were faster to correct solutions (1190.53 ms, $SE = 48.06$) than to incorrect solutions (1357.51 ms, $SE = 55.16$) and were faster to small problems (1121.57 ms, $SE = 43.66$) than to large problems (1438.23 ms, $SE = 62.38$). The advantage for small problems was 317 ms (within-participant $SE = 32$, $d_z = 1.30$), and the advantage for correct solutions was 167 ms (within-participant $SE = 21$, $d_z = 1.05$). Fig. 2 shows the average RT for each solution for the correct solutions only (most solutions appear in only one problem; 12, 18, and 24 each appear in two problems). Fig. 2 illustrates descriptively the effect of solution size on RT, which by definition is the problem size effect as it is operationalized in the literature; statistical analyses to compare individual solutions were not conducted due to the limited number of trials per solution size.

Accuracy

Main effects of correctness, $F(1, 56) = 18.37$, $p < .001$, $\eta^2_G = .07$, and size, $F(1, 56) = 13.87$, $p < .001$, $\eta^2_G = .04$, reached significance, but the interaction between these factors did not, $F(1, 56) = 0.64$, $p = .43$. Mean accuracy was 89.40% (range = 74.11–99.11). Participants were more accurate when responding to correct solutions (91.73%) compared with incorrect solutions (87.08%) and when responding to solutions in small problems (91.09%) compared with large problems (87.71%).

ERP results

ERPs to the solution

N400 latency.: To analyze peak latency, we measured the absolute peak of the N400 correctness effect in the broad window of 200 to 600 ms post-solution in each participant. The mean peak across all channels and participants was 387.87 ms (*SE* = 6.15). In addition, individual participant peak latencies were tested in a repeated-measures ANOVA with 57 participants and within-participant factors of correctness (correct or incorrect), problem size (small or large), and electrode (26 levels). The ANOVA revealed no main effect of correctness, $F(1, 56) = 1.84$, $p = .18$, or size, $F(1, 56) = 1.72$, $p = .20$, and no significant interaction between these factors, $F(1, 56) = 3.01$, p .09. These results are consistent with the N400 component being stable in time (see Kutas & Federmeier, 2011, for a review on the N400).

N400 mean amplitude.: Mean N400 amplitude was measured in a 200-ms window centered around the latency peak (387.87 ms) 288 to 488 ms after solution onset. Fig. 3 shows the grand average ERPs for each condition at each scalp electrode (approximate distribution across the head). Note that the N400 amplitude might fall within the positive range (e.g., across the back of the head) (see Fig. 3). However, the N400 component is measured as a negative-going deflection rather than a negative absolute amplitude. There were significant main effects of correctness, $F(1, 56) = 50.03$, $p < .001$, $\eta^2_G = .04$, and size, $F(1, 56) = 10.31$, $p < .01$, $\eta^2_G = .01$, as well as an interaction between correctness and size, $F(1, 56) = 9.26$, $p < .01$, $\eta^2_G = .01$. The main effect of correctness was due to reduced (less negative) N400s for correct solutions (μV = 4.84, *SE* = 0.47) relative to incorrect solutions (μV = 2.13, *SE* = 0.53). The main effect of size was due to reduced N400s for solutions in small problems (μV = 4.05, *SE* = 0,52) relative to large problems (μV = 2.93, *SE* = 0.46). The interaction was twofold. The main effect of correctness was greater for solutions of small problems, $F(1, 56) = 61.94$, $p < .001$, $\eta^2_G = .16$ (within-participant difference: μV = 3.79, *SE* = 0.48, $d_z = 1.04$), than for solutions of large problems, $F(1, 56) = 8.58$, $p < .01$, $\eta^2_G = .04$ (within-participant difference: μV = 1.64, *SE* = 0.56, $d_z = 0.39$). That is, the effect of correctness was more prominent for smaller problems. Conversely, the effect of size was significant for correct solutions, $F(1, 56) = 20.09$, $p < .001$, $\eta^2_G = .07$ (within-participant difference: μV = 2.20, *SE* = 0.49, $d_z = 0.59$), but did not reach significance for incorrect solutions, $F(1, 56) = 0.01$, $p = .93$ (within-participant difference: μV = 0.05, *SE* = 0.50, $d_z = 0.09$). That is, the main effect of size was driven by correct solutions (see Fig. 4).

To determine the reliability of the N400 problem size effect, additional analyses using standard error (SE) of the mean amplitude were conducted. Using a sampling of $n - 1$ with replacement in R (R Development Core Team, 2016), none of the participants increased the SE of the group when included in the sample. This indicates that the effects were not driven by outliers. The standardized measurement error (SME) tool was then used to quantify single-participant data quality by providing a mean of the standard errors (or SME), where smaller values indicate better data quality (Luck, Stewart, Simmons, & Rhemtulla, 2021). Within the N400 time window (288–488 ms), the average SME value across participants

was 3.81 (range = 1.46–8.69). Correlations between these values and trial counts showed no relationship ($r = .27$, $p = .47$), indicating that even children with lower trial numbers had reliable ERPs. Overall, our data sample is reliable based on these metrics.

The factor of electrode interacted with size, $F(25, 1400) = 3.30$, $p < .01$, $\varepsilon = 0.24$, $\eta^2_G = .003$, and correctness, $F(25, 1400) = 3.41$, $p < .01$, $\varepsilon = 0.21$, $\eta^2_G = .004$, but the three-way interaction among electrode, size, and correctness was not significant, $F(25, 1400) = 0.49$, $p = .83$, $\varepsilon = 0.27$. A distributional ANOVA (described in the "ERP data analysis" section in Method) was run for correctness and size. Size interacted with laterality, $F(1, 56) = 9.32$, $p < .01$, $\eta^2_G = .002$, with larger size effects over medial sites (medial effect = 1.61 μV, $SE = 0.45$; lateral effect = 0.43 μV, $SE = 0.30$). Correctness interacted with laterality, $F(1, 56) = 13.44$, $p < .001$, $\eta^2_G = .002$, and with laterality and anteriority together, $F(3, 168) = 3.17$, $p < .05$, $\varepsilon = 0.86$, $\eta^2_G < .001$, reflecting the typical centromedial distribution of the N400 effect. The N400 effect was larger and more distributed for correctness than for size (see Fig. 5, upper and lower left head plots).

Post hoc analyses were conducted to examine the impact of overall task accuracy on the N400 problem size effect. The range of scores in this sample was 74.1% to 99.1% accuracy. Three performance categories were created by dividing the range of scores into terciles.[5] This resulted in a lower-performing group ($n = 10$; average = 80.2%, range = 74.1–82.4), an average-performing group ($n = 17$; average = 86.4%, range = 82.41–90.8), and a higher-performing group ($n = 30$; average = 94.2%, range = 90.81–99.1). Critically, the majority of children performed above 82% ($n = 47$), which indicates that our sample included children who were good at the task. Although the tercile groups were too small to directly compare their ERP responses, we looked at the brain data after removing the 10 low-performing children. The resulting ERPs ($n = 47$) were similar to the ERPs observed in the whole group ($n = 57$), with small correct solutions eliciting smaller N400 amplitude than large correct and incorrect solutions. The statistical analyses also revealed similar results, namely a main effect of correctness, $F(1, 46) = 42.84$, $p < .001$, a main effect of problem size, $F(1, 46) = 8.85$, $p < .01$, and an interaction between correctness and problem size, $F(1, 46) = 8.08$, $p < .01$. Therefore, the results observed in the whole group did not change when low-performing children were removed, which suggests that (a) our sample was representative of children with overall high math fluency (as measured by accuracy on our task) and (b) even higher-performing children did not show a problem size effect on incorrect solutions.

Post hoc analyses were also conducted to determine the impact of grade level on the problem size effects. There was no significant effect of grade [entered as a between-participant factor, $F(2, 54) = 0.41$, $p = .67$] across mean amplitudes for any of the conditions of interest.

Post-N400 effects.: Directly following the N400 time window, average ERPs per condition were measured over a 300-ms window (488–788 ms post-stimulus) to capture sustained or slow-wave effects. The distribution of the effects is best understood by examining the

[5]Using the median split ($M = 91.1\%$) resulted in a lower-performing group (mean accuracy = 85.0%, range = 74.1–91.1) and a higher-performing group (mean accuracy = 94.4%, range = 92.0–99.1) with drastically different ranges (17% vs. 7%).

topographic plots of the main effects in this time window (Fig. 5), where correctness emerges as a posterior positivity for incorrect solutions (the opposite of the N400 pattern; Fig. 5 Fig. 6) and size emerges as a more central continuation of the prior N400 effect (Fig. 6).

An omnibus ANOVA with factors of correctness (correct or incorrect), size (small or large), and 26 levels of electrode was computed. This analysis obtained a main effect of size, $F(1, 56) = 4.09$, $p < .05$, $\eta^2_G = .004$, which was due to a more positive-going response to solutions of small problems (μV = 5.41, $SE = 0.54$) than large problems (μV = 4.49, $SE = 0.50$). This effect was qualified by an interaction between size and electrode, $F(25, 1400) = 5.15$, $p < .001$, $\varepsilon = 0.25$, $\eta^2_G = .005$. A distributional ANOVA (described in the "ERP data analysis" section in Method) found that size interacted with anteriority, $F(3, 168) = 5.04$, $p < .05$, $\eta^2_G = .003$, and laterality, $F(1, 56) = 10.10$, $p < .05$, $\eta^2_G = .002$, with larger size effects over centromedial channels (effect = 2.74 μV, $SE = 0.71$; see Fig. 5, bottom right). Post hoc tests are reported here with significance of at least p values of .025. An ANOVA with two levels of size (small or large) and 10 levels of electrode (using a cluster of 10 centromedial channels: LMFr, RMFr, LDCe, MiCe, RDCe, LMCe, RMCe, LDPa, MiPa, and RDPa) found that the significant effect of size, $F(1, 56) = 10.69$, $p < .01$, $\eta^2_G = .02$, was due to more positive response to solutions of small problems (μ = 7.65, $SE = .72$) compared with large problems (μ = 5.75, $SE = 0.70$) (within-participant difference: μ = 1.91, $SE = 0.58$, $d_z = 0.43$). This late centromedial effect of problem size may be a continuation of the pattern observed during the earlier N400 time window except as a main effect (no interaction with correctness).

There was no main effect of correctness, $F(1, 56) = 0.02$, $p = .89$, but the interaction between correctness and electrode was significant, $F(25, 1400) = 5.53$, $p < .001$, $\varepsilon = 0.26$, $\eta^2_G = .007$. Distributional analysis (as per above) found that correctness interacted with anteriority, $F(3, 168) = 10.17$, $p < .05$, $\eta^2_G = .07$, and both anteriority and laterality, $F(3, 168) = 3.06$, $p < .05$, $\eta^2_G = .0005$, with larger effects over occipital channels (i.e., one level more posterior than what was found for size) and lateral channels (occipito-lateral effect = 2.16 μV, $SE = 0.60$; see Fig. 3, top right). The effect of correctness was then measured over a cluster of 7 posterior and/or lateral electrodes (LLTe, RLTe, LLOc, LMOc, RMOc, MiOc, and RLOc; see channel labels in Fig. 3) and was examined in an ANOVA with two levels of correctness (correct or incorrect) and 7 levels of electrode, with a corrected p value for significance of .025. This found a main effect of correctness, $F(1, 56) = 10.63$, $p < .001$, $\eta^2_G = .02$, which was due to more positive responses to incorrect solutions (μV = 9.99, $SE = 0.73$) than to correct solutions (μV = 8.19, $SE = 0.66$) (within-participant difference: μV = 1.81, $SE = 0.55$, $d_z = 0.43$), a reversal of the pattern in the N400 time window.

No other effects or interactions were significant in the analysis containing all electrodes (Fs < 1, ps > .10).

ERPs to the second operand—To determine whether problem size effects occurred before the onset of the solution, ERPs were measured from the onset of the second operand (−100 to 900 ms). Visual inspection (see Fig. 7) revealed typical visually evoked potentials (VEPs)—a P1, N1, and P2 over occipital channels and an N1 and P2 over more central and frontal channels—followed by an N400 component that settles back to baseline. There were no visible differences in the ERPs based on problem size. No further analyses are reported.[6]

Additional analyses—A comparison of the behavioral and demographic metrics between the children who were included ($n = 57$) and those who were excluded ($n = 44$)[7] from the ERP analyses was conducted in order to characterize the participants of the current study. Results showed that the two samples did not differ significantly based on age, grade or SES scores (all ps > .15). However, the samples were significantly different on measures of math fluency, $t(98.81) = 3.55$, $p < .001$, accuracy, $t(55.59) = 7.33$, $p < .001$, and response time, $t(99) = -2.44$, $p < .05$, on the task. We summarize the results of the analyses in Table 3. These findings are not surprising given that children with higher math fluency and accuracy necessarily have more trials in each critical condition, making it more likely for them to be included in the ERP analysis of problem size. Not all the excluded children were low performing (lower group $n = 18$, 50%–66%; average group $n = 13$, 66%–82%; higher group $n = 13$, 82%–98%). Although intuitively third graders might be expected to have poorer performance, proportionally more fourth graders were excluded (15/33 third graders, 20/39 fourth graders, and 9/27 fifth graders) from the original sample.

The 44 excluded children showed a main effect of correctness in both accuracy, $F(1, 43) = 9.61$, $p < .005$ (correct = 77% vs. incorrect = 71%), and RT, $F(1, 43) = 45.88$, $p < .005$ (correct = 1427 ms vs. incorrect = 1619 ms). A main effect of problem size was also observed for both accuracy, $F(1, 43) = 49.92$, $p < .001$ (small = 81% vs. large = 67%), and RT, $F(1, 43) = 61.37$, $p < .001$ (small = 1340 ms vs. large = 1706 ms). No significant interaction between these two factors was found for either measure. These results indicate that although the excluded children performed significantly lower than the study sample, they still exhibited a problem size effect similar to the 57 children included in ERP analysis (see “Behavior” section above). This indicates that the included sample is representative of the larger sample from the original study ($n = 99$; Grenier et al., 2020).

Discussion

Overview

This study examined a sizable sample of elementary school children from an existing dataset (Grenier et al., 2020) to characterize how problem size affects access to semantic representations of core factual arithmetic information, namely single-digit multiplication, in the developing mind and brain. ERPs were measured simultaneously with RT and accuracy during a multiplication verification task. The results showed typical effects of behavior, with

[6]An additional analysis was conducted over the baseline (−100 to 0 ms) for the solutions to check for any prolonged effects of problem size after the second operand. There was no effect of problem size, $F(1, 31) = 0.09$, $p = .76$, in this time window, so baseline data did not significantly affect the problem size effect identified at the solution.

[7]Another 2 children were included in addition to the original 99, for a total of 101 children.

lower accuracy and slower RTs for solutions in incorrect and larger problems compared with correct and smaller problems, consistent with the extant behavioral literature (Ashcraft, 1992; Ashcraft & Stazyk, 1981; Campbell & Graham, 1985; Stazyk et al., 1982; Zbrodoff, 1995; Zbrodoff & Logan, 2005). There was no interaction between correctness and problem size, indicating that the effect of size on performance was similar for correct and incorrect solutions (and vice versa).

The N400, an index of semantic memory access, was also compared across correctness and size. As expected, correct solutions elicited smaller N400 amplitude than incorrect solutions, indicating that children attempted to access the solutions from semantic memory, with correct items being facilitated by the preceding operands. When looking at correctness by problem size, the N400 correctness effect (i.e., the difference in amplitude between correct and incorrect solutions) was greater for small problems than for large problems. This suggests that small problems were more facilitated in memory than large problems. In other words, solutions in small problems led to a larger difference between the correct solutions that were strongly expected and the incorrect solutions that were clearly erroneous based on this strong expectation. However, when inverting this interaction to look at problem size by correctness, only the N400 to correct solutions was modulated by problem size, with reduced amplitude for small correct solutions compared with large correct solutions. Notably, this interaction in the ERPs went undetected by behavioral measures. We discuss below the implications of these findings, as well as two effects subsequent to the N400: a main effect of correctness over posterior electrodes and a main effect of problem size over centromedial electrodes.

Effects of problem size

The N400 is a relatively automatic electrophysiological response to any potentially meaningful stimulus regardless of modality (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984). It can be thought of as a window to the current state of semantic memory, a state driven by context and experience. Multiplication facts are learned through memorization in typical American elementary school education (note that differences in education can lead to less reliance on retrieval; Prado et al., 2013), such that the memorized problem (e.g., 2×4) operates as the context for a given solution. As expected in children, the N400 to correct (expected) solutions was smaller in amplitude than that to incorrect (unexpected) solutions (Cerda et al., 2019; Grenier et al., 2020; Moore, Drollette, Scudder, Bharij, & Hillman, 2014; Prieto-Corona et al., 2010). This modulation of the N400 is consistent with facilitation in the memory network and easier processing of correct solutions compared with incorrect solutions. Problem size further modulated the N400 amplitude, but only for correct problems, indicating that only correct solutions are differentially available in memory based on the size of the problem.

The problem size effect is a robust finding in the behavioral literature in children and adults (Ashcraft, 1992; Ashcraft & Stazyk, 1981; Campbell & Graham, 1985; Dickson & Wicha, 2019; Stazyk et al., 1982; Zbrodoff, 1995; Zbrodoff & Logan, 2005). Arithmetic solutions are typically produced (or verified) with less accuracy for larger problems (e.g., 7×8) than for smaller problems (e.g., 2×3). This has been interpreted as evidence that

memory retrieval for larger items is less successful (Campbell & Graham, 1985). Consistent with this, N400 amplitude was reduced for small problems compared with large problems. However, breaking this effect down by correctness reveals more nuance in the underlying cognitive processes.

Problems that exist in memory: Correct solutions

As mentioned in the introduction, the problem size effect observed in behavior has been explained by three interrelated factors: frequency of occurrence, interference or "confusion" in memory, and strategy. Both frequency and interference are known to modulate the N400 to individual words, with larger amplitude for words that are less frequent or that have greater competition (Fischer-Baum et al., 2014; Meade et al., 2018; Megías & Macizo, 2016; Van Petten, 1993; Rugg, 1990). Small problems are more frequently encountered than large problems (Ashcraft & Christy, 1995; Geary, 1996) and may also experience less interference (or confusion) than large problems (for a review, see Ashcraft & Guillaume, 2009), making both frequency and competition possible sources of the N400 problem size effect. That is, the difference between small and large problems could reflect either greater facilitation in memory for more frequent small problems, greater interference in memory for large problems, or some combination of both. We discuss below, when unpacking the results for incorrect solutions, how frequency may be the more parsimonious explanation.

With regard to strategy, deployment of alternate strategies to direct retrieval, such as repeated addition and transformation, is more frequently reported for large problems than for small problems (cf. Kirk & Ashcraft, 2001; LeFevre, Sadesky, et al., 1996; Siegler, 1988). In principle, children might have used different strategies on a trial-by-trial basis[8] to arrive at the solution before it appeared, although evidence from the literature does suggest that children quickly adopt a retrieval strategy when learning multiplication facts (Siegler, 1988).

Even if the problem size effect is partially driven by the deployment of procedural strategies (e.g., counting, transforming) for large problems, it is unlikely that this explains the observed N400 pattern. First, there is no evidence in the literature that strategy per se modulates N400 amplitude. Indeed, strategy implies a conscious process, yet the N400 is an automatic brain response to meaningful or potentially meaningful information and can occur even without conscious awareness of a stimulus (Luck et al., 1996). Simply put, the fact that both small and large problems elicit an N400 indicates an attempt to access both types of problems from semantic memory. Critically, if children were engaging slower procedural methods to respond to large problems, a delay in processing the solutions for meaning would be expected. Yet, even though the operands and solution are presented at the same speed for both small and large problems, the N400 did not shift in time based on problem size.

It might be challenging to square the data from consciously reported strategy with an explanation of the observed N400 modulation that does not acknowledge strategy. However, the N400 is an index of only one early step in the cascade of cognitive events before children press a button to make their correctness judgment. Consciously reported strategy

[8]Trial-by-trial analysis is not feasible with ERPs in this study because it would have required many more trials than what children would tolerate.

effects are more likely reflected during stages of processing other than the initial attempt to access meaning reflected in the N400. As noted, adults elicit a P300 on this same paradigm, reflecting more superficial processing of correct solutions such as targets in a categorization task (Polich, 1987, 2007, 2012; Sutton et al., 1965, 1982). Given that children do not show the adult P300 response, not even the small solutions are being processed as over-learned targets as in adults. We discuss problem size effects that occur after the N400 below.

In brief, both small and large problems elicit an N400, reflecting engagement of semantic processes for both problem types. Considering the broader literature on the N400, small problems likely generated a well-constrained semantic context for the correct (expected) solutions (see Jost, Hennighausen, & Rösler, 2004, for further discussion). In other words, at the level of semantic memory, children were more prepared to process correct solutions when they appeared in small problems than when they appeared in large problems. This would be consistent with the idea that the N400 reflects the current memory state as driven by available context. Future research can test more explicitly whether conscious strategy can modulate the N400 in children or the P300 in adults for simple arithmetic tasks or more broadly.

Problems that do not exist in memory: Incorrect solutions

The incorrect solutions in this study were always a multiple of one of the operands and therefore a possible correct solution on related “times table” problems. Theoretically, then, if children recall multiplication facts in a network of related facts, the N400 for the incorrect solution might have been modulated by the relative predictability of the correct solution. More specifically, with more facilitation for small problems, there would be more spread of activation to related solutions, and therefore small incorrect related solutions would also elicit smaller (more facilitated) N400 amplitude than large incorrect problems. In contrast, the N400 to incorrect solutions did not differ based on problem size.

If the N400 problem size effect could be explained by the differential deployment of strategy across small and large problems, we might expect that smaller problems would be more easily retrieved from memory and that large problems would be more laboriously calculated. In the language literature, there is some evidence that tasks that demand more superficial or automatic processing of words can reduce typical modulations of the N400 such as the lexical frequency effect, where less frequent words elicit larger N400 amplitude (Fischer-Baum et al., 2014). By analogy, items that require a more procedural strategy should elicit larger N400 effects, but this is not what we see for large problems that elicited a smaller effect of correctness.

This result also poses a challenge to confusion or competition at the level of semantics as an explanation for differences in speed and/or accuracy based on problem size when rejecting incorrect solutions (Ashcraft & Guillaume, 2009). As discussed above, under an interference account, larger problems may have a greater history of retrieval errors during learning and in turn may be more likely to activate wrong answers. This competition would have elicited larger amplitude for large incorrect solutions than for small incorrect solutions, but instead there was no difference.

The null effect also mitigates potential concern that numerical distance could account for the reported problem size effect in behavior (i.e., faster to reject more distant incorrect solutions than closer incorrect solutions). Related incorrect solutions are inherently further in numerical distance from the correct solution in large problems than in small problems (here, ±4 vs. ± 7 on average). For example, solutions in the 3 × table are + 3 from each other (3, 6, 9, …), but solutions in the 9 × table are + 9 from each other (9, 18, 27, …). The larger distance from the correct solution in large problems should have made them easier to dismiss than smaller closer solutions (Ashcraft & Stazyk, 1981; De Smedt, Verschaffel, & Ghesquière, 2009; Niedeggen & Rösler, 1999; see De Smedt, Noël, Gilmore, & Ansari, 2013, for a review). Behaviorally, children were not faster or more accurate at verifying large versus small incorrect problems (i.e., no interaction between size and correctness). The N400 theoretically should have been reduced in amplitude for easier to dismiss items (more distant solutions) if numerical distance were an important factor, but again there was no difference between small (close) and large (distant) incorrect solutions.

Overall, frequency of encountering small versus large problems is a more parsimonious explanation of the results for both correct and incorrect solutions than confusion or strategy. However, frequency is likely not the sole cognitive factor contributing to the ease of accessing multiplication facts from memory. Fig. 2 shows that RTs increase with increasing solution size. It is unlikely that each increment in solution size would incrementally be less likely to occur (i.e., less frequent). In the broader numerical cognition literature, studying effects of problem size in arithmetic might be taken as an attempt to capture sensitivity to numerical magnitude itself (Siegler & Braithwaite, 2017). So, perhaps this incremental change in RT reflects a sensitivity to the magnitude of the solution itself (which operationally is the definition of problem size here). Importantly, by accessing the magnitude of the solutions, children are accessing the “meaning” of the solutions as they verify the multiplication facts, which would support our hypothesis that the N400 elicited in children reflects access to semantic memory.

In the current study, however, numerosity was not deliberately isolated as a factor in the design. The correct solutions exist on a continuum of small to large and were dichotomized into size categories, as is typical in the arithmetic literature (whereas studies of numerosity typically generate distinct classes of magnitude for comparison). Therefore, numerosity is confounded with other factors such as frequency of occurrence. Most critically, there was no effect of problem size on the N400 to incorrect solutions, where the magnitude of the solutions should have also had an effect.

Even in the behavioral findings, magnitude of the solution does not completely explain the problem size effect. In Fig. 2, solutions in the 10-s increments (i.e., 10, 20, 30, 40) appear to elicit faster RTs than the adjacent solutions. Although this was not statistically measured due to small trial counts per solution, this pattern would be consistent with the previously reported “five effect” (Campbell & Graham, 1985; Lemaire & Reder, 1999; Masse & Lemaire, 2001; Siegler, 1988). Multiplication problems with 5 as an operand have correct products that end in 0 or 5 (e.g., less variability in the final digit of the solutions compared with other problems). These items are potentially learned using a different strategy than other problems (Siegler, 1988). Once learned, these “special” items are easier to retrieve

from semantic memory. As far as how this might affect the N400, we predict that these "special" problems that are easier to retrieve from memory would lead to a reduction in N400 amplitude when verifying those facts. This would be most consistent with the leading explanation of the N400 as an index of semantic memory processes.

Effects of problem size beyond the N400

Importantly, these results affirm that the N400 and overt behavioral responses can be independent measures of cognition (see similar argument in Federmeier & Kutas, 1999). Generally speaking, the N400 indexes an early stage of semantic processing. Sometimes N400 effects can occur in the absence of behavioral effects (McLaughlin et al., 2004), whereas other times there are behavioral effects with no modulation of N400 (Heinze, Muente, & Kutas, 1998). Here, the interaction between problem size and correctness on the N400 was not measurable in either RT or accuracy, both of which showed only main effects of problem size.

The divergence between brain and behavior measures reveals that the source of the behavioral problem size effect for incorrect problems is not at the initial stage of semantic processing indexed by the N400. Instead, problem size likely affected the processing of incorrect solutions at a later cognitive stage, perhaps during reevaluation of the accurately rejected solutions or difficulty in response selection/execution (e.g., inhibiting responses to large problems due to uncertainty, second-guessing disproportionately for large problems).

Indeed, directly following the N400 time window, there was a main effect of problem size manifested as a sustained difference through the end of the recording epoch, with small problems eliciting more positive amplitude than large problems independent of correctness (see Fig. 3). This effect over-lapped in scalp distribution with the N400 (Fig. 5, bottom plots), so it is possible that it reflects a slow return to baseline or a continuation of the N400 problem size effect. However, problem size also modulated the response to incorrect solutions, unlike in the earlier N400 window.

The cognitive significance of this effect is unclear, but it is consistent with the view that problem size continues to affect processing beyond initial access to semantic memory. This later downstream effect may be the type of processing that leads to intertrial confusion and more repetition effects observed in other experimental designs (e.g., retrieval-induced forgetting) (Galfano et al., 2011; Phenix & Campbell, 2004).

Do children show an adult-like brain response?

A second post-N400 effect occurred as a main effect of correctness over occipital electrodes. This effect was an inversion of the N400 effect, with incorrect solutions eliciting larger positive amplitude than correct solutions independent of problem size (Fig. 5, top plots). We consider two possible explanations for this finding.

In the sentence comprehension literature, the N400 can be followed by a late positive component (LPC) that emerges with semantic incongruities and is thought to reflect post-N400 reprocessing (Coulson & Kutas, 2001; Kuperberg, 2007). In the adult arithmetic literature, a similar positivity has also been described as an LPC (Niedeggen et al., 1999;

Núñez-Peña, 2008; Szucs & Csépe, 2005). When children encountered incorrect solutions, they may have engaged additional processing mechanisms to further assess them (cf. Prieto-Corona et al., 2010), double-checking the problem in memory or second guessing their knowledge. Therefore, it is possible that this late positivity reflects a general semantic reprocessing LPC.

However, children showed this effect over occipital electrodes, which is more posterior than the typical distribution for an LPC but perhaps similar to a posteriorly distributed P300. Therefore, it is possible that this positivity reflects the beginning of an adult-like P300, albeit delayed. Notably, in the larger sample from the original Grenier et al. (2020) study, this posterior late positivity for correctness did not reach significance. As mentioned in Method, children who met the threshold for inclusion here were typically better at the verification task than children who were excluded. Therefore, this later positivity, which only reached significance in this smaller sample, may be sensitive to multiplication fluency, reflecting the beginning of an adult-like P300 in more fluent children.

Adults do not exhibit a modulation on the N400 (Grenier et al., 2020) and correspondingly do not have differing levels of semantic preactivation for solutions as a function of problem size (Dickson & Federmeier, 2017; Dickson & Wicha, 2019). In addition, adults show a problem size effect on both correct and incorrect solutions (Dickson & Wicha, 2019), whereas children show a problem size effect on correct solutions only. These outcomes support a gradual transition, from engaging semantic memory processes in children (N400) to more direct solution categorization in adults (P300). To be clear, it is not that adults have formed a different kind of memory but rather that they can access the information more efficiently with less depth of processing than children.

In children, correct trials are likely more sensitive to problem size because their accessibility in semantic memory is affected by relative frequency of exposure, a property that does not differ across incorrect solutions. As children switch to the adult approach of treating the problems as a target categorization task, both correct and incorrect problems are affected by problem size, with larger problems being harder to categorize in both cases. Interestingly, even the high-performing children in our sample did not show a problem size effect on incorrect solutions, suggesting that the full transition to adult-like processing happens beyond fifth grade.

This transition might not be an abrupt shift but rather a gradual change; that is, children might begin to use target categorization to solve small problems but still require semantic access to verify large problems. This idea is in line with the overlapping waves theory (Siegler, 1996), which also states that different strategies remain available over development but that the frequency of use changes over time, with more efficient strategies (here target categorization) becoming more dominant. As strategy use changes throughout development, children become less reliant on hippocampal memory (long-term memory) and transition into engaging more automatic processes (i.e., target categorization) to verify arithmetic facts (Smith & Squire, 2009).

These results speak to the broader understanding of meaning in arithmetic in a couple of ways. The idea that the N400 reflects access to meaning creates an interesting theoretical dilemma. Models of arithmetic cognition propose that the "meaning" of arithmetic facts is located in the magnitude code, which in theory resides in the parietal cortex (Campbell & Clark, 1992; Dehaene & Cohen, 1995). However, the N400 is thought to be generated in verbally mediated areas of temporal cortex (Lau, Phillips, & Poeppel, 2008). This arithmetic N400 in children either reflects a "meaning" representation outside of the parietal cortex, which would be inconsistent with current models of arithmetic, or this is a novel parietal-generated N400 (which cannot be determined with ERP data alone). In either case, the evidence that children elicit an N400 for both small and large problems is an important revelation.

Verification versus production

Finally, this study had children verify presented solutions, which may not entail the same processes or yield the same outcomes as a task that requires children to produce the solutions (e.g., Ashcraft, Fierman, & Bartolotta, 1984; Campbell & Tarling, 1996; Zbrodoff & Logan, 1990). In the memory domain verification and production are analogous to recognition versus recall memory (Craik & McDowd, 1987; Haist, Shimamura, & Squire, 1992; Rawson & Zamary, 2019; Rhodes, Greene, & Naveh-Benjamin, 2019), and in the language domain they are analogous to comprehension versus production (Glenberg & Gallese, 2012; Pickering & Garrod, 2013; Rommers, Dell, & Benjamin, 2020). Thus, it is not unreasonable to expect differences.

We argue that the memory system itself is the same across tasks. What changes is more likely the degree of processing (i.e., when and how the information is being accessed and extracted). We hypothesize that in a production task children would rely more on semantic memory to retrieve the solution. Interestingly, in the language domain, questions of *whether* readers preactivate upcoming words (i.e., mentally produce upcoming words during comprehension) have been replaced by questions of *when* and *how* readers preactivate upcoming words (Delong, Troyer, & Kutas, 2014; Federmeier, Kutas, & Schul, 2010; Kuperberg & Jaeger, 2016; Wlotko & Federmeier, 2015). We believe that a similar outcome could be found in arithmetic verification tasks based on manipulations of answer types, task instructions, timing, and other factors yet to be determined.

We can make a speculative comparison of the ERP findings with data from the only known arithmetic production task in children (Van Beek et al., 2014, 2015). In these studies, children spoke the solutions to simple addition problems that were presented all at once (e.g., 2 + 3). A problem size effect was observed in ERPs measured to the onset of the problem, with more negative amplitude for large problems than for small problems on anterior electrodes between 250 and 500 ms and on posterior electrodes between 500 and 625 ms. The closest comparison point in our task would be to look at the ERPs from the onset of the second operand when children could begin to solve the problem (Fig. 7). In our data, problem size did not modulate the ERPs at the second operand, indicating that children were not treating the problems differently at that point in time. These findings may support the long-standing argument that production and verification tasks engage

different cognitive processes (Campbell & Tarling, 1996; Zbrodoff & Logan, 1990) and that verification can be done using alternate rules or patterns (e.g., Krueger, 1986). However, the multiple differences between the methods used (addition vs. multiplication, presenting the whole problem vs. one operand at a time in sequence, etc.) make it hard to draw this conclusion definitively. Future research could directly compare production and verification on identical tasks to determine whether the cognitive processes engaged in each are indeed different before the solution is presented.

Conclusion

In brief, problem size only affects semantic level processing in children for problems that are available in memory (i.e., correct solutions), as revealed by an N400 modulation for correct solutions only. We argue that the frequency of encountering small versus large correct solutions may be the most parsimonious account of this problem size effect on correct solutions. A late effect of problem size for incorrect solutions reveals that the observed behavioral problem size effects may be due to later categorization or decision-related processes, not differences in access to semantic memory. This highlights the value of synchronously assessing behavior and brain indices of cognition. This study adds to the current math cognition literature by characterizing the quality of memory for single-digit multiplication tables in the developing brain and provides support for a gradual transition from accessing multiplication facts from semantic memory in children to engaging more automated target categorization in adults. Future research could investigate individual differences in the use of semantic memory versus rote memory across age and populations.

Acknowledgments

This work received computational support from the University of Texas at San Antonio's HPC Cluster Shamu, operated by the Office of Information Technology. We thank all the participants and their families for their participation. The study was supported by grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R21HD079884 and R21HD098878) and by a BRAIN EAGER grant (1451032) from the National Science Foundation. The project was funded in part by the University of Texas at San Antonio, Office of the Vice President for Research, Economic Development, and Knowledge Enterprise.

References

Ashcraft MH (1982). The development of mental arithmetic: A chronometric approach. Developmental Review, 2, 213–236.

Ashcraft MH (1992). Cognitive arithmetic: A review of data and theory. Cognition, 44, 75–106. [PubMed: 1511587]

Ashcraft MH, & Christy KS (1995). The frequency of arithmetic facts in elementary texts: Addition and multiplication in Grades 1–6. Journal for Research in Mathematics Education, 26, 396–421.

Ashcraft MH, Fierman BA, & Bartolotta R (1984). The production and verification tasks in mental addition: An empirical comparison. Developmental Review, 4, 157–170.

Ashcraft MH, & Guillaume MM (2009). Mathematical cognition and the problem size effect. In Ross B (Ed.). Psychology of learning and motivation (Advances in Research and Theory) (Vol. 51, pp. 121–151). Amsterdam: Elsevier.

Ashcraft MH, & Stazyk EH (1981). Mental addition: A test of three verification models. Memory & Cognition, 9, 185–196. [PubMed: 7242333]

Bakeman R (2005). Recommended effect size statistics for repeated measures designs. Behavior Research Methods, 37, 379–384. [PubMed: 16405133]

Butterworth B, Reeve R, Reynolds F, & Lloyd D (2008). Numerical thought with and without words: Evidence from indigenous Australian children. Proceedings of the National Academy of Sciences of the United States of America, 105, 13179–13184. [PubMed: 18757729]

Campbell JID (1987). Production, verification, and priming of multiplication facts. Memory & Cognition, 15, 349–364. [PubMed: 3670055]

Campbell JID, & Austin S (2002). Effects of response time deadlines on adults' strategy choices for simple addition. Memory & Cognition, 30, 988–994. [PubMed: 12450100]

Campbell JID, & Clark JM (1992). Cognitive number processing: An encoding-complex perspective. In Campbell JID (Ed.). The nature and origins of mathematical skills (Advances in Psychology) (Vol. 91, pp. 457–491). Amsterdam: Elsevier.

Campbell JID, & Graham DJ (1985). Mental multiplication skill: Structure, process, and acquisition. Canadian Journal of Psychology/Revue Canadienne de Psychologie, 39, 338–366.

Campbell JID, & Tarling DM (1996). Retrieval processes in arithmetic production and verification. Memory & Cognition, 24, 156–172. [PubMed: 8881320]

Carrillo-de-la-Peña M, & Cadaveira F (2000). The effect of motivational instructions on P300 amplitude. Neurophysiologie Clinique/Clinical Neurophysiology, 30, 232–239. [PubMed: 11013896]

Cerda VR, Grenier AE, & Wicha NYY (2019). Bilingual children access multiplication facts from semantic memory equivalently across languages: Evidence from the N400. Brain and Language, 198, 104679. [PubMed: 31445417]

Coulson S, & Kutas M (2001). Getting it: Human event-related brain response to jokes in good and poor comprehenders. Neuroscience Letters, 316, 71–74. [PubMed: 11742718]

Craik FI, & McDowd JM (1987). Age differences in recall and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 474–479.

De Brauwer J, Verguts T, & Fias W (2006). The representation of multiplication facts: Developmental changes in the problem size, five, and tie effects. Journal of Experimental Child Psychology, 94, 43–56. [PubMed: 16376370]

De Smedt B, Noël MP, Gilmore C, & Ansari D (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. Trends in Neuroscience and Education, 2, 48–55.

De Smedt B, Verschaffel L, & Ghesquière P (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. Journal of Experimental Child Psychology, 103, 469–479. [PubMed: 19285682]

Dehaene S, & Cohen L (1995). Towards an anatomical and functional model of number processing. Mathematical Cognition, 1, 83–120.

Delong KA, Troyer M, & Kutas M (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. Language and Linguistics Compass, 8, 631–645. [PubMed: 27525035]

Delorme A, & Makeig S (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods, 134, 9–21. [PubMed: 15102499]

Dickson DS, Cerda VR, Beavers RN, Ruiz A, Castañeda R, & Wicha NYY (2018). When 2 × 4 is meaningful: The N400 and P300 reveal operand format effects in multiplication verification. Psychophysiology, 55 e13212 [PubMed: 30132910]

Dickson DS, & Federmeier KD (2017). The language of arithmetic across the hemispheres: An event-related potential investigation. Brain Research, 1662, 46–56. [PubMed: 28237544]

Dickson DS, & Wicha NYY (2019). P300 amplitude and latency reflect arithmetic skill: An ERP study of the problem size effect. Biological Psychology, 148, 107745. [PubMed: 31470071]

Domahs F, & Delazer M (2005). Some assumptions and facts about arithmetic facts. Psychology Science, 47, 96–111.

Federmeier KD (2022). Connecting and considering: Electrophysiology provides insights into comprehension. Psychophysiology, 59, e13940. [PubMed: 34520568]

Federmeier KD, & Kutas M (1999). A rose by any other name: Long-term memory structure and sentence processing. Journal of Memory and Language, 41, 469–495.

Federmeier KD, & Kutas M (2001). Meaning and modality: Influences of context, semantic memory organization, and perceptual predictability on picture processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 202–224.

Federmeier KD, Kutas M, & Schul R (2010). Age-related and individual differences in the use of prediction during language comprehension. Brain and Language, 115, 149–161. [PubMed: 20728207]

Federmeier KD, Wlotko EW, De Ochoa-Dewald E, & Kutas M (2007). Multiple effects of sentential constraint on word processing. Brain Research, 1146, 75–84. [PubMed: 16901469]

Fischer-Baum S, Dickson DS, & Federmeier KD (2014). Frequency and regularity effects in reading are task dependent: Evidence from ERPs. Language, Cognition and Neuroscience, 29, 1342–1355.

Frank MC, Everett DL, Fedorenko E, & Gibson E (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. Cognition, 108, 819–824. [PubMed: 18547557]

Galfano G, Penolazzi B, Fardo F, Dhooge E, Angrilli A, & Umiltà C (2011). Neurophysiological markers of retrieval-induced forgetting in multiplication fact retrieval. Psychophysiology, 48, 1681–1691. [PubMed: 21824154]

Geary DC (1996). The problem-size effect in mental addition: Developmental and cross-national trends. Mathematical Cognition, 2, 63–94.

Geary DC, Hoard MK, & Bailey DH (2012). Fact retrieval deficits in low achieving children and children with mathematical learning disability. Journal of Learning Disabilities, 45, 291–307. [PubMed: 21252374]

Gelman R, & Gallistel CR (2004). Language and the origin of numerical concepts. Science, 306, 441–443. [PubMed: 15486289]

Glenberg AM, & Gallese V (2012). Action-based language: A theory of language acquisition, comprehension, and production. Cortex, 48, 905–922. [PubMed: 21601842]

Grenier AE, Dickson DS, Sparks CS, & Wicha NYY (2020). Meaning to multiply: Electrophysiological evidence that children and adults treat multiplication facts differently. Developmental Cognitive Neuroscience, 46, 100873. [PubMed: 33129033]

Haist F, Shimamura AP, & Squire LR (1992). On the relationship between recall and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 691–702.

Heinze HJ, Muente TF, & Kutas M (1998). Context effects in a category verification task as assessed by event-related brain potential (ERP) measures. Biological Psychology, 47, 121–135. [PubMed: 9554184]

Hollingshead A (1975). Four factor index of social status. Yale University. Unpublished manuscript.

Imbo I, & Vandierendonck A (2008). Effects of problem size, operation, and working-memory span on simple-arithmetic strategies: Differences between children and adults? Psychological Research, 72, 331–346. [PubMed: 17457605]

Isreal JB, Chesney GL, Wickens CD, & Donchin E (1980). P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. Psychophysiology, 17, 259–273. [PubMed: 7384376]

Jasinski EC, & Coch D (2012). ERPs across arithmetic operations in a delayed answer verification task. Psychophysiology, 49, 943–958. [PubMed: 22563982]

Jost K, Hennighausen E, & Rösler F (2004). Comparing arithmetic and semantic fact retrieval: Effects of problem size and sentence constraint on event-related brain potentials. Psychophysiology, 41, 46–59. [PubMed: 14693000]

Jung TP, Makeig S, Humphries C, Lee TW, McKeown MJ, Iragui V, & Sejnowski TJ (2000). Removing electroencephalographic artifacts by blind source separation. Psychophysiology, 37, 163–178. [PubMed: 10731767]

Kirk EP, & Ashcraft MH (2001). Telling stories: The perils and promise of using verbal reports to study math strategies. Journal of Experimental Psychology: Learning, Memory, and Cognition. 10.1037/0278-7393.27.1.157.

Kleih SC, Nijboer F, Halder S, & Kübler A (2010). Motivation modulates the P300 amplitude during brain-computer interface use. Clinical Neurophysiology. 10.1016/j.clinph.2010.01.034.

Koshmider JW, & Ashcraft MH (1991). The development of children's mental multiplication skills. Journal of Experimental Child Psychology. 10.1016/0022-0965(91)90077-6.

Krueger LE (1986). Why 2×2=5 looks so wrong: On the odd-even rule in product verification. Memory & Cognition, 27, 157–175.

Kuperberg GR (2007). Neural mechanisms of language comprehension: Challenges to syntax. Brain Research, 1146, 23–49. [PubMed: 17400197]

Kuperberg GR, & Jaeger TF (2016). What do we mean by prediction in language comprehension? Language, Cognition and Neuroscience, 31, 32–59.

Kutas M, & Federmeier KD (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology, 62, 621–647.

Kutas M, & Hillyard SA (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature, 307, 161–163. [PubMed: 6690995]

Lau EF, Phillips C, & Poeppel D (2008). A cortical network for semantics: (De)constructing the N400. Nature Reviews Neuroscience, 9, 920–933. [PubMed: 19020511]

Lawrence M (2016). Package “ez.” https://github.com/mike-lawrence/ez.

LeFevre JA, Daley KE, Buffone L, Greenham SL, Bisanz J, & Sadesky GS (1996). Multiple routes to solution of single-digit multiplication problems. Journal of Experimental Psychology: General, 125, 284–306.

LeFevre JA, Sadesky GS, & Bisanz J (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 216–230.

Lemaire P, & Reder L (1999). What affects strategy selection in arithmetic? The example of parity and five effects on product verification. Memory & Cognition, 27, 364–382. [PubMed: 10226446]

Lemaire P, & Siegler RS (1995). Four aspects of strategic change: Contributions to children’s learning of multiplication. Journal of Experimental Psychology: General, 124, 83–97. [PubMed: 7897342]

Lopez-Calderon J, & Luck SJ (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience, 8. 10.3389/fnhum.2014.00213.

Luck SJ, Vogel EK, & Shapiro KL (1996). Word meanings can be accessed but not reported during the attentional blink. Nature, 383, 616–618. [PubMed: 8857535]

Luck S, Stewart A, Simmons A, & Rhemtulla M (2021). Standardized measurement error: A universal measure of data quality for averaged event-related potentials. Psychophysiology, 58, e13793. [PubMed: 33782996]

Masse C, & Lemaire P (2001). Do people combine the parity- and five-rule checking strategies in product verification? Psychological Research, 65, 28–33. [PubMed: 11505610]

McLaughlin J, Osterhout L, & Kim A (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. Nature Neuroscience, 7, 703–704. [PubMed: 15195094]

Meade G, Grainger J, Midgley KJ, Emmorey K, & Holcomb PJ (2018). From sublexical facilitation to lexical competition: ERP effects of masked neighbor priming. Brain Research, 1685, 29–41. [PubMed: 29407530]

Megías P, & Macizo P (2016). Simple arithmetic: Electrophysiological evidence of coactivation and selection of arithmetic facts. Experimental Brain Research, 234, 3305–3319. [PubMed: 27423447]

Miller K, Perlmutter M, & Keating D (1984). Cognitive arithmetic: Comparison of operations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 46–60.

Moore RD, Drollette ES, Scudder MR, Bharij A, & Hillman CH (2014). The influence of cardiorespiratory fitness on strategic, behavioral, and electrophysiological indices of arithmetic cognition in preadolescent children. Frontiers in Human Neuroscience, 8. 10.3389/fnhum.2014.00258.

Niedeggen M, & Rösler F (1999). N400 effects reflect activation spread during retrieval of arithmetic facts. Psychological Science, 10, 271–276.

Niedeggen M, Rösler F, & Jost K (1999). Processing of incongruous mental calculation problems: Evidence for an arithmetic N400 effect. Psychophysiology, 36, 307–324. [PubMed: 10352554]

Noël MP, Fias W, & Brysbaert M (1997). About the influence of the presentation format on arithmetical-fact retrieval processes. Cognition, 63, 335–374. [PubMed: 9265874]

Núñez-Peña MI (2008). Effects of training on the arithmetic problem-size effect: An event-related potential study. Experimental Brain Research, 190, 105–110. [PubMed: 18648782]

Núñez-Peña MI, & Escera C (2007). An event-related brain potential study of the arithmetic split effect. International Journal of Psychophysiology, 64, 165–173. [PubMed: 17360062]

Peters L, & De Smedt B (2018). Arithmetic in the developing brain: A review of brain imaging studies. Developmental Cognitive Neuroscience, 30, 265–279. [PubMed: 28566139]

Phenix TL, & Campbell JID (2004). Effects of multiplication practice on product verification: Integrated structures model or retrieval-induced forgetting? Memory & Cognition, 32, 324–335. [PubMed: 15190723]

Pickering MJ, & Garrod S (2013). How tightly are production and comprehension interwoven? Frontiers in Psychology, 4. 10.3389/fpsyg.2013.00238.

Polich J (1987). Task difficulty, probability, and inter-stimulus interval as determinants of P300 from auditory stimuli. Electroencephalography and Clinical Neurophysiology: Evoked Potentials, 68, 311–320. [PubMed: 2439311]

Polich J (2007). Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology, 118, 2128–2148. [PubMed: 17573239]

Polich J (2012). Neuropsychology of P300. In Luck SJ & Kappenman ES (Eds.), The Oxford handbook of event-related potential components (pp. 159–188). New York: Oxford University Press.

Prado J, Lu J, Liu L, Dong Q, Zhou X, & Booth JR (2013). The neural bases of the multiplication problem-size effect across countries. Frontiers in Human Neuroscience, 7. 10.3389/fnhum.2013.00189.

Prieto-Corona B, Rodríguez-Camacho M, Silva-Pereyra J, Marosi E, Fernández T, & Guerrero V (2010). Event-related potentials findings differ between children and adults during arithmetic-fact retrieval. Neuroscience Letters, 468, 220–224. [PubMed: 19897015]

R Development Core Team (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rawson KA, & Zamary A (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. Journal of Memory and Language, 105, 141–152.

Rhodes S, Greene NR, & Naveh-Benjamin M (2019). Age-related differences in recall and recognition: A meta-analysis. Psychonomic Bulletin & Review, 26, 1529–1547. [PubMed: 31396816]

Rommers J, Dell GS, & Benjamin AS (2020). Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory. Cognition, 198, 104206. [PubMed: 32035323]

Rugg MD (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. Memory & Cognition, 18, 367–379. [PubMed: 2381316]

Siegler RS (1988). Strategy choice procedures and the development of multiplication skill. Journal of Experimental Psychology: General, 117, 258–275. [PubMed: 2971762]

Siegler RS (1996). Emerging minds: The process of change in children's thinking. New York: Oxford University Press.

Siegler RS, & Braithwaite DW (2017). Numerical development. Annual Review of Psychology, 68, 187–213.

Smith CN, & Squire LR (2009). Medial temporal lobe activity during retrieval of semantic memory is related to the age of the memory. Journal of Neuroscience, 29, 930–938. [PubMed: 19176802]

Stazyk EH, Ashcraft MH, & Hamann MS (1982). A network approach to mental multiplication. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 320–335.

Sutton S, Braren M, Zubin J, & John ER (1965). Evoked-potential correlates of stimulus uncertainty. Science, 150, 1187–1188. [PubMed: 5852977]

Sutton S, Ruchkin DS, Munson R, Kietzman ML, & Hammer M (1982). Event-related potentials in a two-interval forced-choice detection task. Perception & Psychophysics, 32, 360–374. [PubMed: 7155782]

Szucs D, & Csépe V (2005). The effect of numerical distance and stimulus probability on ERP components elicited by numerical incongruencies in mental addition. Cognitive Brain Research, 22, 289–300. [PubMed: 15653300]

Van Beek L, Ghesquière P, De Smedt B, & Lagae L (2014). The arithmetic problem size effect in children: An event-related potential study. Frontiers in Human Neuroscience, 8. 10.3389/fnhum.2014.00756.

Van Beek L, Ghesquière P, De Smedt B, & Lagae L (2015). Arithmetic difficulties in children with mild traumatic brain injury at the subacute stage of recovery. Developmental Medicine and Child Neurology, 57, 1042–1048. [PubMed: 26268837]

Van Petten C (1993). A comparison of lexical and sentence-level context effects in event-related potentials. Language and Cognitive Processes, 8, 485–531.

Wechsler D (2009). Wechsler Individual Achievement Test (3rd ed.). San Antonio, TX: Psychological Corporation.

Wicha NY, Dickson DS, & Martinez-Lincoln A (2018). Arithmetic in the bilingual brain. In Berch DB, Geary DC, & Koepke KM (Eds.), Language and culture in mathematical cognition (pp. 145–172). San Diego: Elsevier Academic Press.

Wlotko EW, & Federmeier KD (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. Cortex, 68, 20–32. [PubMed: 25987437]

Woodcock RW, McGrew KS, & Mather N (2001). Woodcock-Johnson III Tests of Cognitive Abilities. Itasca, IL: Riverside.

Zbrodoff NJ (1995). Why is 9 + 7 harder than 2 + 3? Strength and interference as explanations of the problem-size effect. Memory & Cognition, 23, 689–700. [PubMed: 8538442]

Zbrodoff NJ, & Logan GD (1990). On the relation between production and verification tasks in the psychology of simple arithmetic. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 83–97.

Zbrodoff NJ, & Logan GD (2005). What everyone finds: The problem-size effect. In Campbell JID (Ed.), Handbook of mathematical cognition (pp. 331–345). New York: Psychology Press.

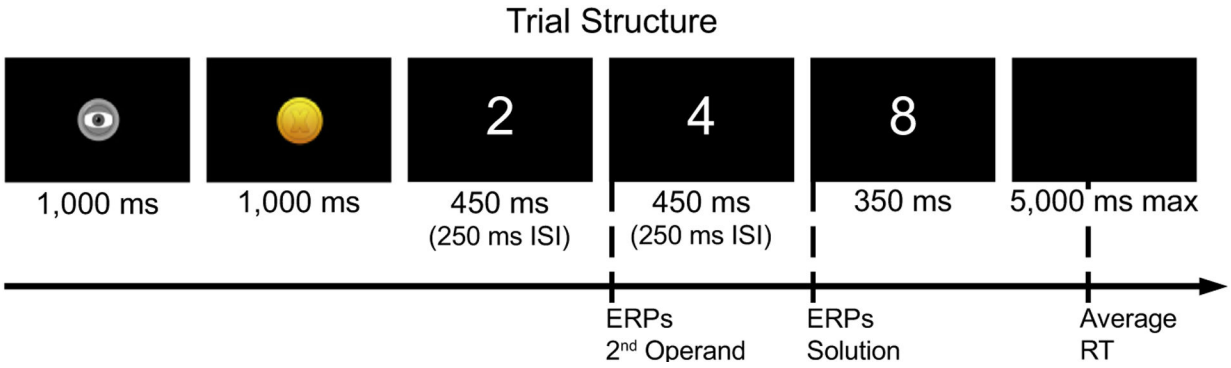

Fig. 1.

An example of a multiplication problem showing the trial structure from left to right over time (in milliseconds). Horizontal dashed lines indicate the time-locking points for critical event-related brain potentials (ERPs) as well as the average response time (RT) (1274 ms). ISI, interstimulus interval.

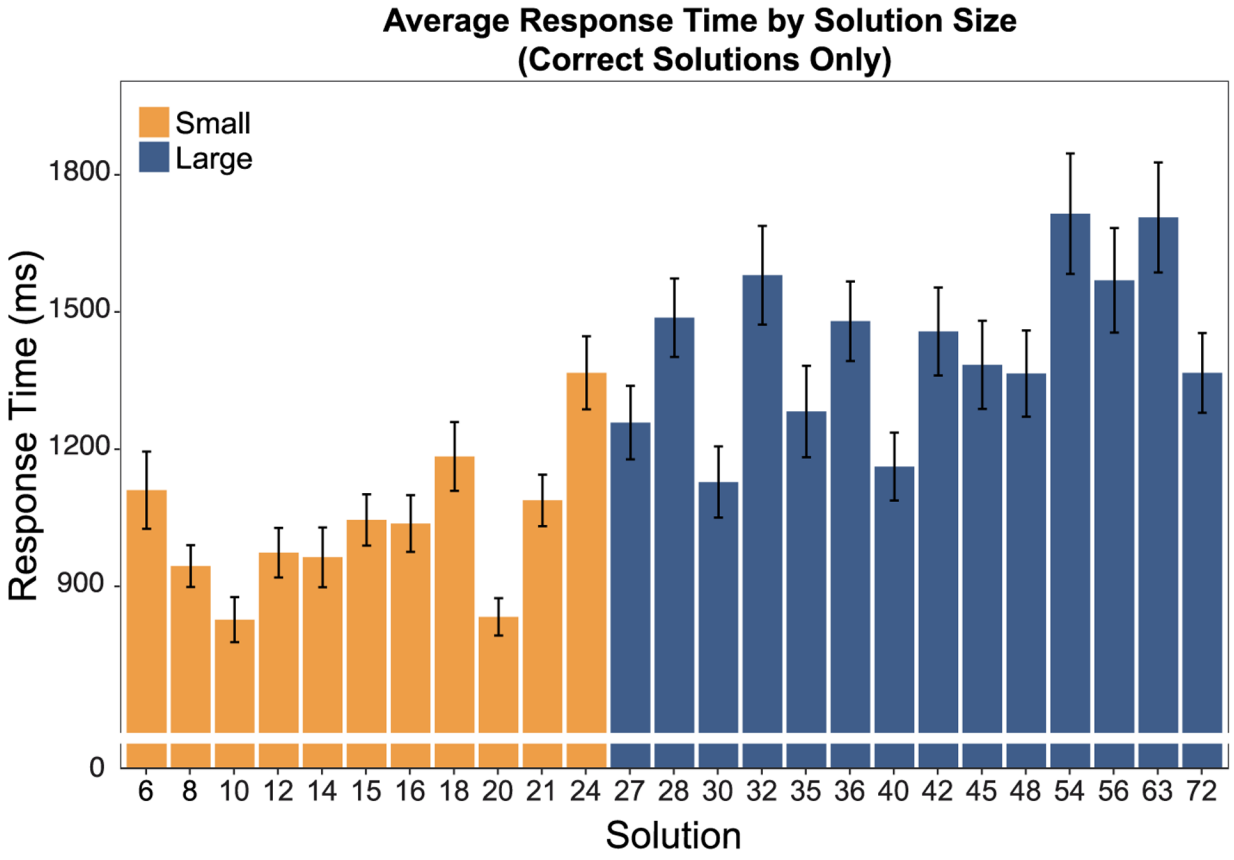

Fig. 2.

Average response times by solution size for correct trials only. Small problems (in orange; left side) elicit faster response times than larger problems (in blue; right side). Individual solutions were not compared statistically.

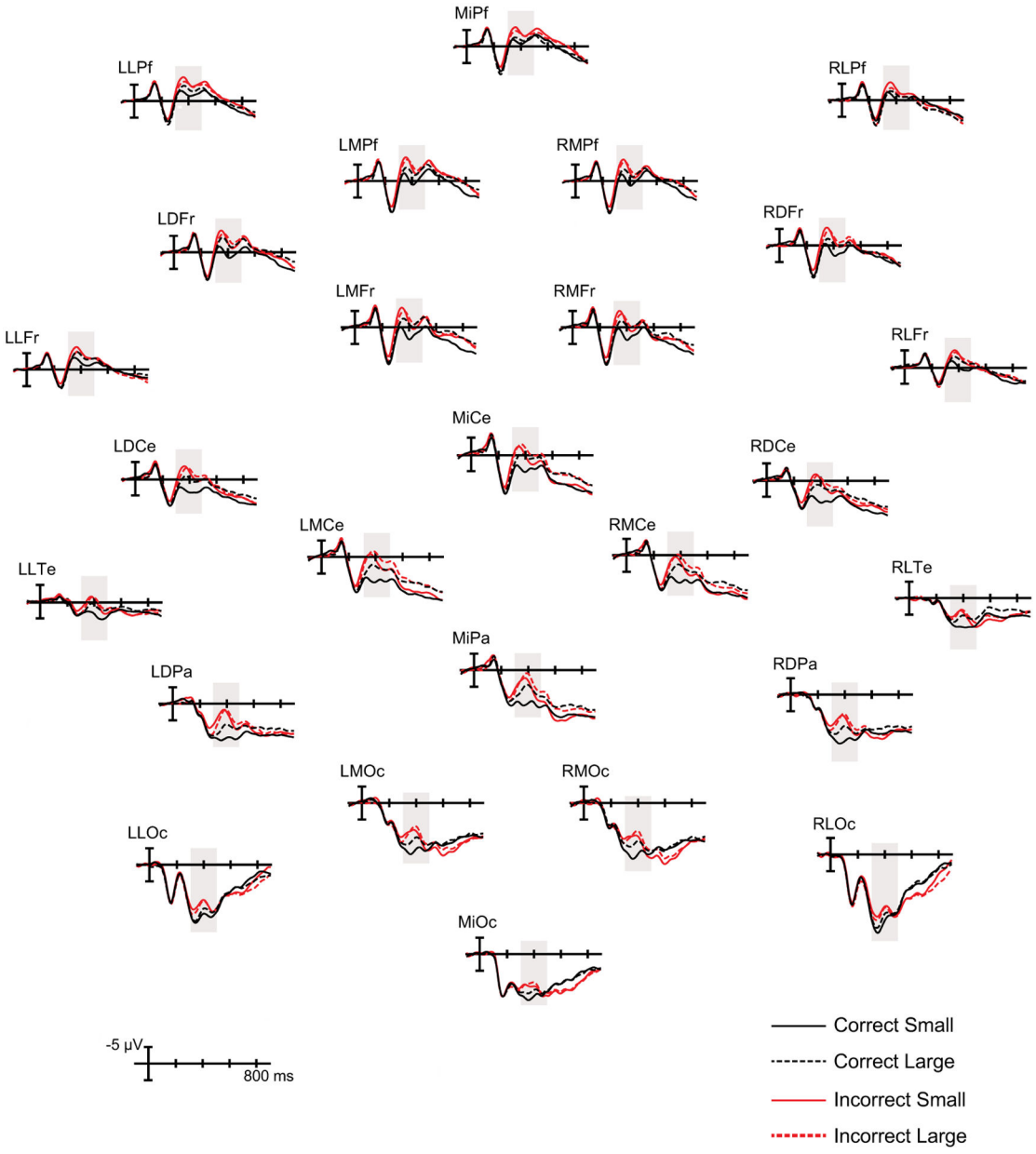

Fig. 3.

Grand average event-related brain potentials to solutions are plotted for the 26 scalp electrodes used in the analysis, with front channels at the top (labeled according to scalp location). Time (in milliseconds) is on the x axis, and voltage (in microvolts) is on the *y* axis, with negative voltage plotted up. Correct solutions (in black) elicit less negative N400s compared with incorrect solutions (in red). The effect of size on the N400 is prominent on correct solutions, shown by the difference between solid and dashed lines.

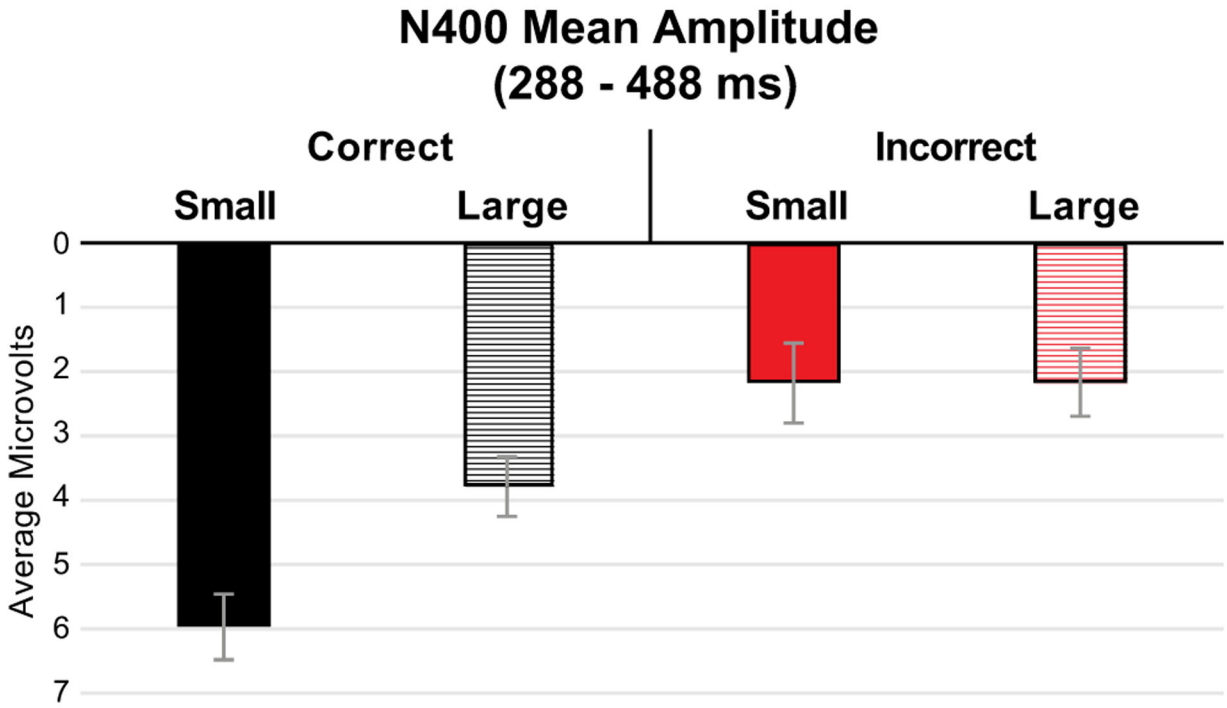

Fig. 4.

Interaction between correctness and problem size. Mean amplitude of the N400 (in microvolts; 288–488 ms poststimulus) averaged across all electrodes for each condition is plotted with standard error ($n = 57$). Positive is plotted down to be consistent with the ERP plots.

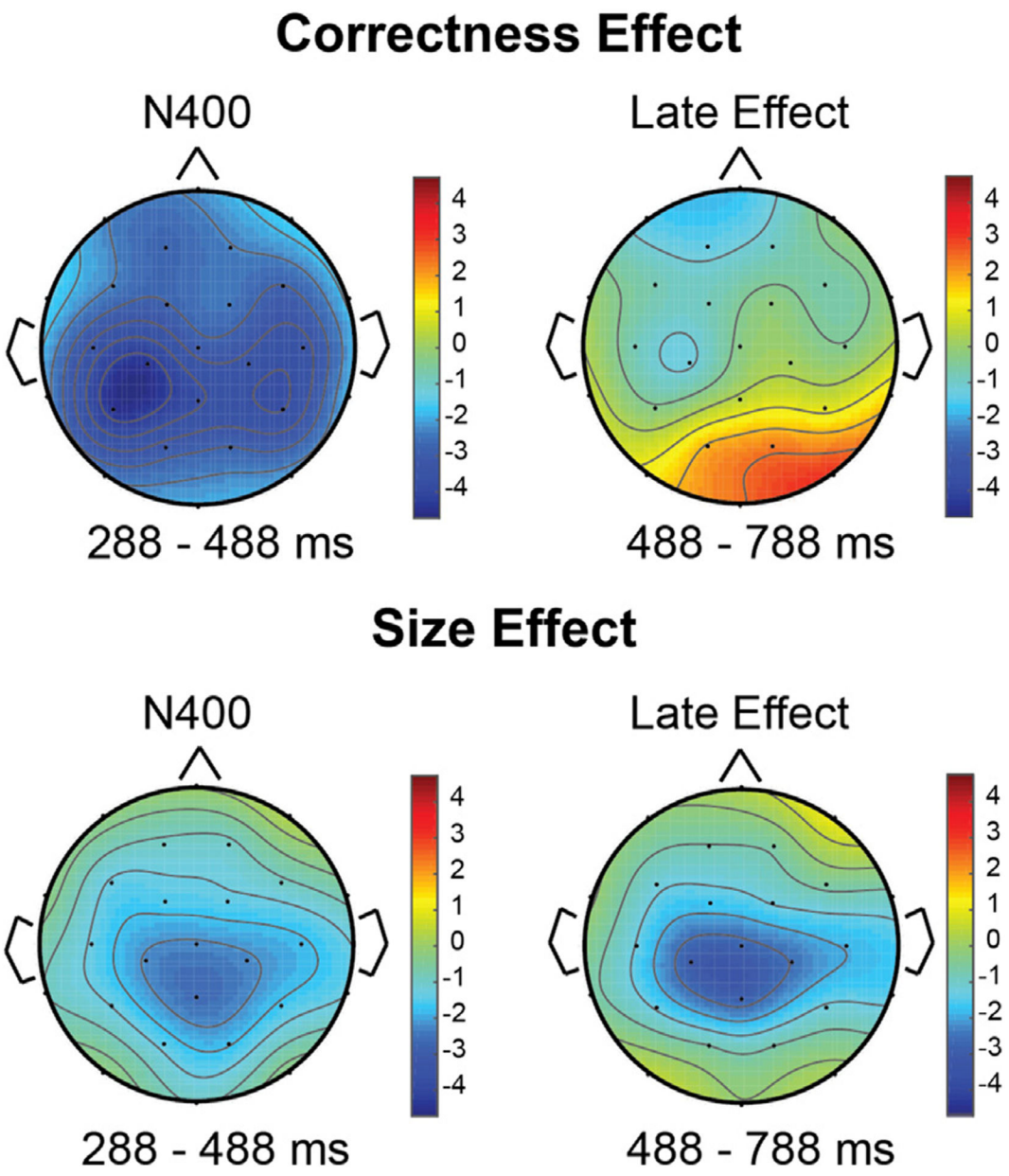

Fig. 5.

Topographical plots of the main effects of correctness (top; incorrect minus correct) and size (bottom; large minus small) for both the N400 time window (left) and post-N400 time window (right). The scales are matched across plots to facilitate comparison. The correctness effect changes polarity and location after the N400 time window, whereas the size effect does not.

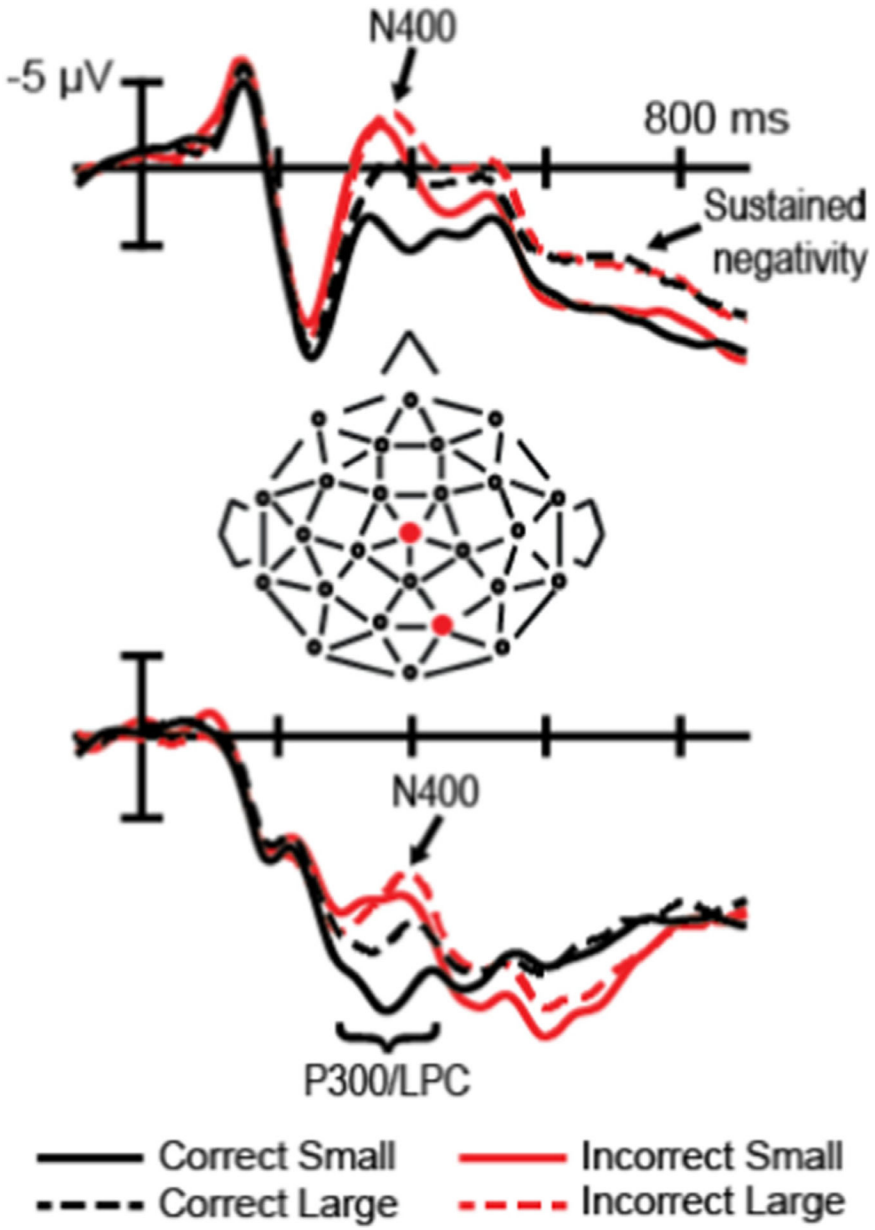

Fig. 6.

Grand average event-related brain potentials for each condition plotted from solution onset for two representative electrodes from Fig. 3 (MiCe and RMOc), as indicated on the scalp map with red dots. The effects of interest are labeled on the two plots. LPC, late positive component.

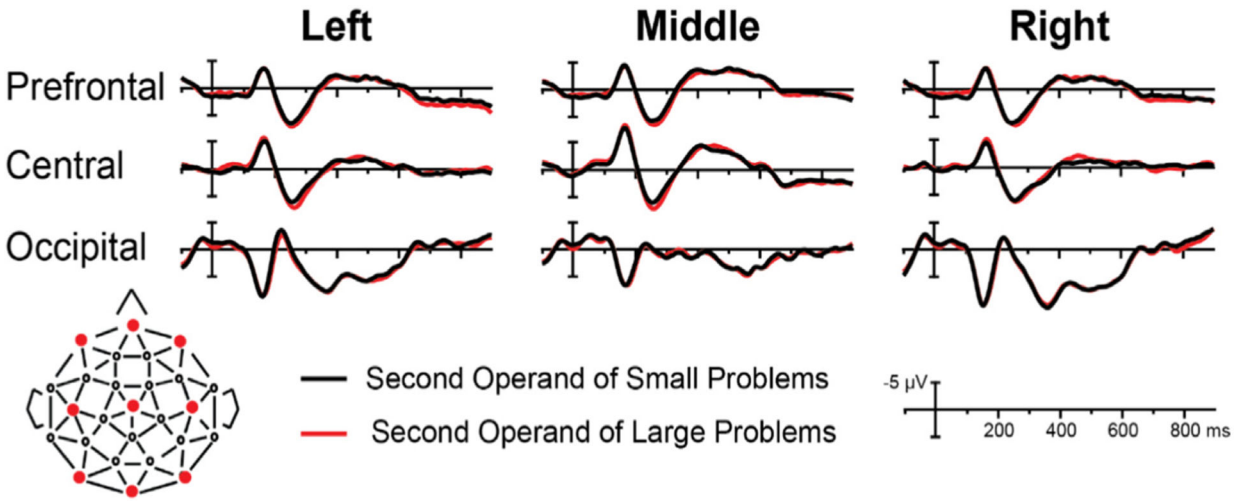

Fig. 7.

Grand average event-related brain potentials time-locked to the onset of the second operand. The six representative electrodes are shown with small (black) and large (red) problems overlapping across the epoch.

Table 1

Distribution of the number of participants included in the averaged event-related brain potentials per trial count ranges for each condition.

Trial count	Correct small	Correct large	Incorrect small	Incorrect large
15–19	11	17	15	24
20–23	23	28	26	27
24–28	24	13	17[a]	7[a]

[a]There were no participants with 28 trials for these conditions.

Table 2

Offline performance on standardized cognitive measures (presented for completeness and comparison with original sample from Grenier et al., 2020).

Assessment	Children (*n* = 57)	
	Mean (*SE*)	**Range**
Math fluency[a]		
Addition	106.00 (1.75)	72–143
Subtraction	110.33 (1.49)	93–152
Multiplication	113.82 (1.90)	90–159
Working memory[b]	110.12 (2.04)	74–157
Phonological awareness[c]	102.21 (1.84)	70–131
Vocabulary size[d]	100.33 (1.35)	76–124
Oral comprehension[e]	108.07 (1.61)	83–134

Note. The means are standardized scores where 100 is the grade-based norm and 15 points reflects 1 standard deviation outside the norm.

[a]Math fluency was measured by the math fluency task of the Wechsler Individual Achievement Test - Third Edition (Wechsler, 2009).

[b]Working memory was measured by the numbers reversed task of the Woodcock–Johnson III Tests of Achievement (WJ-III).

[c]Phonological awareness was measured by the incomplete words task of the WJ-III.

[d]Vocabulary size was measured by the picture vocabulary task of the WJ-III.

[e]Oral comprehension was measured by the oral comprehension task of the WJ-III.

Table 3

Comparison of the means (and standard errors) of demographic and cognitive measures for children included in the current study and those excluded from the original sample of Grenier et al. (2020).

	Included children (*n* = 57)	**Excluded children (*n* = 44)**	***t* Test result**
Grade	4.69 (0.10)	4.51 (0.11)	$t(99) = 1.17$, $p = .25$
Age (years)	10.07 (0.11)	9.90 (0.14)	$t(99) = 1.10$, $p = .27$
Socioeconomic status	49.84 (1.53)	46.59 (1.79)	$t(96) = 1.38$, $p = .17$
Math fluency (standardized score)	113.82 (1.90)	104.43 (1.74)	$t(98.81) = 3.55$, $p < .001$
Accuracy (%)	90.12 (0.79)	73.86 (2.07)	$t(55.59) = 7.33$, $p < .001$
Response time (ms)	1279.28 (51.49)	1490.63 (72.77)	$t(99) = -2.44$, $p < .05$