Published in final edited form as:

Psychophysiology. 2018 November ; 55(11): e13212. doi:10.1111/psyp.13212.

When “2 × 4” Is Meaningful: The N400 and P300 Reveal Operand Format Effects in Multiplication Verification

Danielle S. Dickson^{a,*}, Vanessa R. Cerda^a, Rosemary N. Beavers^c, Andres Ruiz^d, Ricardo Castañeda^a, and Nicole Y.Y. Wicha^{a,b,e}

^aDepartment of Biology, The University of Texas at San Antonio

^bUTSA Neurosciences Institute, The University of Texas at San Antonio

^cThe University of Texas Medical Branch, School of Medicine

^dTexas Tech University Health Sciences Center El Paso

^eResearch Imaging Institute, UT Health San Antonio

Abstract

Arithmetic problems share many surface-level features with typical sentences. They assert information about the world, and readers can evaluate this information for sensibility by consulting their memories as the statement unfolds. When people encounter the solution to the problem 3×4 , the brain elicits a robust event-related potential effect as a function of answer expectancy (12 being the expected completion; 15 being unexpected). Initially, this was labeled an N400 effect, implying that semantic memory had been accessed. Subsequent work suggested instead that the effect was driven by a target P300 to the correct solutions. The current study manipulates operand format to differentially promote access to language-based semantic representations of arithmetic. Operands were presented either as spoken number words or as sequential Arabic numerals. The critical solution was always an Arabic numeral. In Experiment 1, the correctness of solutions preceded by spoken operands modulated N400 amplitude, whereas solutions preceded by Arabic numerals elicited a P300 for correct problems. In Experiment 2, using only spoken operands, the delay between the second operand and the Arabic numeral solution was manipulated to determine if additional processing time would result in a P300. With a longer delay, an earlier N400 and no distinct P300 was observed. In brief, highly familiar digit operands promoted target detection, whereas spoken numbers promoted semantic level processing – even when solution format itself was held constant. This provides evidence that the brain can process arithmetic fact information at different levels of representational meaningfulness as a function of symbolic format.

Keywords

arithmetic; spoken number words; numerical cognition; ERPs; N400; P300

*Corresponding author: 1 UTSA Circle Dr. Department of Biology, University of Texas, San Antonio, San Antonio, TX 78249, danielle.s.dickson@gmail.com.

1. Introduction

The sentence “I gave each of my 2 friends 3 apples for a total of 6 apples” can be expressed succinctly as “two times three equals six”. This expression can still be read as a semantically and grammatically correct sentence containing nouns (the numbers), a preposition (times) and a verb (equals), but the meaning conveyed becomes an abstract manipulation of quantities, which is difficult to visualize absent the concrete descriptors of *apples* and *friends*. When the number words are replaced with Arabic digits and mathematical symbols (“ $2 \times 3 = 6$ ”), the sentence appears even further removed from the typical domain of meaningful language. However, there is some evidence that the brain may indeed be treating this expression like it would a simple sentence (Jost, Hennighausen, & Rösler, 2004; Niedeggen, Rösler, & Jost, 1999). Namely, despite behavioral differences in response times and accuracy (Bernardo, 2002; Frenck-Mestre & Vaid, 1993), several electrophysiological studies have reported similar brain responses to arithmetic problems presented as number words and as Arabic digits (Bassok, Fisher, & Osterhout, 2009; Fisher, Bassok, & Osterhout, 2010; Niedeggen, & Rösler, 1999; Niedeggen et al., 1999; Salillas & Wicha, 2012).

These electrophysiological findings were interpreted in support of the view that number words and Arabic digits may each access conceptual representations of arithmetic, and that processing arithmetic expressions for meaning proceeds similarly across format types (Bassok et al., 2009). Some have argued that multiplication tables and other simple arithmetic facts are stored as linguistic representations in verbal memory, which are then consulted during answer retrieval even if the problems are written in Arabic digit format (Dehaene, 1992). However, recent studies have looked again at these effects and discovered that arithmetic problems presented as Arabic digits in fact do show a different response than what is seen in traditional language paradigms (Dickson & Federmeier, 2017; Jasinski & Coch, 2012).

Here, we used event-related potentials (ERPs) to test how number format impacts the cognitive processes recruited during assessment of multiplication problems. We record responses to Arabic digit solutions preceded by either Arabic digit numerals or by spoken number words, with the idea that these formats might differentially encourage semantic processing of the arithmetic. Our results have implications for the leading models of math cognition, as we will discuss below. We begin with a review of prior studies to construct more specific expectations of possible ERP responses.

1.1 ERP Evidence for Arithmetic Processing

The first studies using ERPs to study arithmetic expressions reported language-like patterns of brain activity, namely an N400-like effect, to the solution for simple multiplication problems (Niedeggen et al., 1999). The N400, a negative-going voltage shift peaking around 400ms after the onset of a stimulus, was first observed in response to semantically incongruous sentences (Kutas & Hillyard, 1980), and has since been observed in response to any meaningful, or potentially meaningful, stimulus (see Kutas & Federmeier, 2011 for a review). N400 sensitivity to semantics empowered language researchers to address long-standing questions in linguistics (e.g., about the relationship between syntactic and semantic processing) by recording the response to words (often sentence-final) that varied in

expectancy given a prior sentential context. Compared to a congruent or expected continuation to a sentence, the amplitude of the N400 increases as the provided continuation is increasingly further from what would be expected based on the preceding context (Kutas & Hillyard, 1984).

The sentence-like structure of simple arithmetic expressions permits analogous measurement of ERPs to their final stimulus, the solution, in an arithmetic sentence. In fact, the neural response observed by Niedeggen and colleagues (1999) was reminiscent of this classic N400 effect. Incorrect solutions appeared to elicit a larger negative-going amplitude compared to correct solutions, with the effect peaking around 400ms after the solution was presented. This result was readily accepted into the broader N400 literature because it was compatible with the common view that the N400 component is responsive to meaningful information in any modality (Kutas & Federmeier, 2011). Thus, it was assumed, when the sentence-like expression 2×3 was completed with an incongruous number, like 7, the larger N400 amplitude reflected the *amodal* expectation for a different (correct) number based on the *meaning* of the preceding context.

This N400 effect to answers presented as Arabic digits was striking because it implied that symbolic arithmetic expressions elicit expectations for the solution at the level of semantics (Jost et al., 2004; Niedeggen & Rösler, 1999). This interpretation is consistent with dominant views of how arithmetic facts are stored and accessed from memory (e.g., Campbell, 1987a,b; Dehaene, 1992). Similar to semantic memory networks for language, mathematical facts show priming effects and production errors that reflect a memory network of closely related math facts (e.g., Campbell, 1987a,b). Number word representations in verbal memory may even be consulted during retrieval of memorized arithmetic facts regardless of the presentation format (Dehaene, 1992), which would similarly predict language-like N400 effects. However, this interpretation of the arithmetic ERP effect has come under scrutiny in recent years.

1.2 Current Interpretations of Arithmetic ERP Effects

Obtaining an N400 effect for an unexpected item is not always the outcome for brain responses to meaningful patterns and sequences. For example, deviant notes in simple familiar melodies do not systematically modulate N400 amplitudes because the amount of conceptual information conveyed by the melodic context is impoverished (Besson & Macar, 1987; Paller, McCarthy & Wood, 1992). Yet, with some encouragement to tap into richer aspects of the meaning of musical elements (e.g., cross-modal priming studies with language stimuli), it is possible to obtain N400 modulations using these types of stimuli (Daltrozzo & Schön, 2009; Koelsch et al., 2004). Similarly, even a simple patch of color can generate an N400 effect provided that the preceding context leads readers to predict a specific color (e.g., Brang, Edwards, Ramachandran, & Coulson, 2008), and abstract novel line drawings can modulate N400 amplitudes if participants are encouraged to think about their potential meaningfulness (i.e., if a sketch is interpreted as a poorly drawn animal) (Voss, Schendan, & Paller, 2010). Thus, Arabic digit and number word solutions should also show N400 modulations – provided readers are encouraged to interpret them as meaningful entities.

After more than a decade of ERP research into the topic, it has become clear that the brain responses to different types of proposed arithmetic solutions are more than just simple N400 effects. Numerical expressions presented in number words do generate sentence-like N400 responses with larger negative amplitude for incorrect than correct solutions (e.g., Salillas & Wicha, 2012). However, numerical expressions presented in Arabic digits elicit a difference driven primarily (if not solely) by a prominent *positive* amplitude modulation for *correct* solutions, rather than a modulation of the negative-going N400 (Dickson & Federmeier, 2017; Jasinski & Coch, 2012). In fact, most experiments using digits, and even some using number words, appear to feature the positivity in response to correct solutions (Bassok et al., 2009; Fisher et al., 2010; Guthormsen et al., 2016; Martinez-Lincoln, Cortinas, & Wicha, 2015; Salillas & Wicha, 2012; Szűcs & Csépe, 2004), but this positivity has not been reported as such in most of these studies. Avancini, Soltész, and Szűcs (2015) acknowledge the presence of positivities for both correct and incorrect solutions, but oddly argue that a latency shift between them resulted in the emergence of the “N400-like” response. Critically, the early positivity for correct responses may be the primary contributor to both the amplitude and the *timing* of the reported arithmetic effect.

The first explicit report of this positivity was in Jasinski and Coch (2012), where they identified it as a target P300. The target P300 is a positive-going amplitude modulation peaking around 300ms, with variable latency dependent on task demands and largest amplitude when a participant confidently detects a rare and desired target stimulus (Polich, 2012). Under this interpretation, participants treat the correct solution for a multiplication problem as a salient *target* in a categorical judgment task – the typical eliciting condition for a P300 (Johnson & Donchin, 1980). This view is sensible when one considers that the participants, who were college students, came to the experiment with years of experience retrieving answers to simple arithmetic problems and verifying that their answers were correct. This also may help explain why the arithmetic ERP effect appears earlier than a typical N400. Whereas a typical semantic N400 peaks around 400ms poststimulus, the so-called arithmetic N400 effect occurs up to 100ms earlier, as Fisher and colleagues (2010) noted, consistent in timing with a P300.

Ironically, in the historical context of ERP component discovery, the original sentence processing study that discovered the N400 component was expecting to find a P300 for expected sentential completions, and instead had to explain its surprising absence (Kutas & Hillyard, 1980). In the domain of math, the component identification process has occurred in reverse: there was an assumption that there *should* be an N400 effect when comparing expected solutions to unexpected solutions, and thus there has been slow recognition that there is instead a large P300 effect, especially in the Arabic digit format. The reason this effect can be interpreted as two very different brain responses is because the absolute value of ERP amplitude is not, in and of itself, easily interpretable. Instead, an *effect* is defined by how the ERP waveform in one condition compares to the ERP waveform in another condition. In an ideal design, the comparison condition would be a neutral control stimulus, but, as in most psychological research of complex behavior, rarely does such a beast exist. Instead, researchers typically assume that a congruent condition should elicit the “normal” brain response, and any deviation in amplitude from this pseudo-baseline captures and reflects the effect of interest. In the case of arithmetic problems, the assumption was that the

incorrect solutions were the deviants, and attention was thus focused on their relative negativity.

This theoretically-driven bias to see an N400 was emphasized by showing the effect as a difference ERP with incorrect solutions (less positive) minus correct solutions (more positive) (e.g. Fig. 3, Niedeggen & Rösler, 1999). Mathematically, this leads to a negative amplitude difference (e.g., $(+1) - (+3) = -2$) supporting the interpretation of the difference as driven by a negative shift in amplitude, rather than a positivity. Of course, the same difference subtracted in reverse (e.g., $(+3) - (+1) = +2$) would be more easily interpretable as a positive deflection for the correct compared to incorrect solutions. Another reason, perhaps, for interpreting the arithmetic effect as an N400 is that incorrect solutions do appear to modulate the N400 modulation under certain circumstances, in addition to or in place of the P300 to correct responses (Galfano, Mazza, Angrilli, & Umiltà, 2004; Galfano, Penolazzi, Vervaeck, Angrilli, & Umiltà, 2009; Salillas & Wicha, 2012; Szűcs & Soltész, 2010). In fact, whether the effect derived from judging the correctness of solutions is composed of a P300 and/or an N400 seems to depend on the format of the multiplication problems¹.

1.3 The Current Study

Thus far, there has been ERP evidence for meaningfulness processing (i.e., N400 effects) in arithmetic when number words are used as stimuli (with some ambiguity regarding the overlap with the P300) (Bassok et al., 2009; Fisher et al., 2010; Guthormsen et al., 2016; Martinez-Lincoln et al., 2015; Salillas & Wicha, 2012). In contrast, arithmetic problems presented as Arabic digits result in much smaller, if not absent, N400 effects (Jasinski & Coch, 2012). Attempts to uncover implicit processing of meaningfulness during incidental arithmetic comprehension with digits has led only to small and atypical-looking N400-like effects (Galfano et al., 2004, 2009). Thus far, the format of the context and the format of the solution have been conflated (cf., Szűcs & Csépe, 2004). The solution itself might not be the critical factor driving the presence or absence of the P300 (or the presence of the N400). Instead, the preceding context might be sufficient to elicit these different effects, reflecting differences in the preparatory processes for upcoming answer assessment.

In the current study, we manipulate the format of the context (operands) and keep the format of the solution the same to determine if the context can modulate the degree of meaningfulness, as indexed by the N400, in processing simple multiplication problems. To this end, in Experiment 1 we record the response to Arabic numeral solutions to simple arithmetic problems (e.g., “24”), preceded by the same operands presented either as spoken number words (“six times four”) or visual Arabic numbers (“ 6×4 ”). By looking at the ERPs to exactly the same solutions preceded by different contextual formats, we will both determine if arithmetic representations are sensitive to the format of the problem

¹Keeping in mind that any meaningful stimulus can elicit an N400, and that the amplitude of the N400 is smallest for contextually congruent, expected stimuli, it is possible that correct solutions presented as Arabic digits elicit an N400 effect that is too small to be independently identified from the much larger co-occurring P300 effect (as suggested by Galfano et al., 2004, 2009). However, although the P300 can obscure and overlap in time with the N400, in cases when it has not overlapped, there was no second, isolated modulation of the N400. For example, Dickson and Federmeier (2017) showed a correctness effect that peaked at 330ms and resolved prior to the typical time window for an N400 effect, and yet no subsequent N400 effect was observed.

(independent of the format of the answer), and potentially identify (in a within-subject design) a factor that can drive the generation of different brain responses to arithmetic solutions. Based on previous findings, Arabic digit solutions should elicit a P300 when preceded by digits and an N400 when preceded by words, reflecting differences in contextual semantic support, rather than an inability to link the solutions to meaning. In Experiment 2, we determine if increasing the processing delay between the second operand and the solution affects the ERPs observed to number words.

2. EXPERIMENT 1

2.1 Method

2.1.1 Participants—Twenty-six participants were included for analysis in this study. All participants were right-handed, native monolingual English speakers (7 male) with normal or corrected-to-normal vision and none had cognitive or neurological impairments (including history of brain injury) or current use of psychoactive medications. Participants' average age was 21 (range: 18 - 28), with an average of 0.82 (range: 0.47-1) on the Edinburgh handedness inventory (where +1 indicates strongly right-handed and -1 strongly left-handed; Oldfield, 1971). Informed consent was received in accordance with the Institutional Review Board (IRB) at the University of Texas at San Antonio.

Participants were recruited from the local University population, had at least a high school education level, and reported their highest level of mathematics coursework as calculus or below (i.e., math majors were excluded). Due to the high proportion of bilingual speakers in the local language environment, special efforts were made to ensure that participants had not experienced substantial exposure to any language other than English² (through the Language Experience and Proficiency Questionnaire or LEAP-Q; Marian, Blumenfeld, & Kaushanskaya, 2007). All participants reported no more than 4 years of foreign language courses (standard in U.S. education), had not spent more than 2 weeks in a non-English speaking country, and self-reported no greater than low level proficiency (self-rated 2 on a scale of 7) in a foreign language. No participants reported living in a home environment with individuals who spoke a language other than English. A sequence of rising and falling pure tones was used to ensure that the sounds were intelligible at a comfortable sound level for the task.

2.1.2 Offline behavioral cognitive assessments—To characterize cognitive abilities related to math ability, we obtained measures of working memory, phonological processing, and math fluency (results summarized in Table 1). Phonological awareness was assessed through a composite score from the Elision and Blended Words subtests of the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999). Working memory was assessed through a subtest of the Woodcock Johnson Test of Cognitive Abilities (Numbers Reversed, WJ III COG; Woodcock, McGrew, & Mather, 2001b, normative scores derived from WJ III NU, Woodcock, McGrew, Mather & Schrank, 2007). Math fluency was measured through a subtest of the Woodcock Johnson Tests of

²Additional language proficiency measures were also collected for future comparison to a bilingual sample, the results of which are outside the scope of the current monolingual report and are not reported here.

Achievement in which participants solved a mixture of single-digit arithmetic problems (addition, subtraction, and multiplication operations intermixed) with a 3-minute time limit (Math Fluency, WJ III ACH, Woodcock, McGrew, & Mather, 2001a; normative scores also derived from WJ III NU).

After EEG collection was complete, each participant also reported their strategy for solving multiplication problems through an adapted version of Campbell and Xue's (2001) strategy questionnaire. Ten multiplication problems were presented with operands in Arabic digit format, and ten additional multiplication problems were presented with operands in auditory number word format. After each problem, participants identified what primary strategy they had just employed (options: transformation, counting, remembering, other).

2.1.3 Stimuli—We chose to use auditory number words (instead of written words) primarily because we are motivated by a broader desire to reduce the potential confound of reading fluency in our measures of arithmetic processing. Although auditory stimuli have already been successfully used in other sentence processing studies (e.g., Federmeier, McLennan, De Ochoa, & Kutas, 2002), they have not been previously used to test sensitivity to meaningfulness in the arithmetic domain. Thus, establishing a functional auditory paradigm through which to examine arithmetic processing independent of reading ability forms a secondary goal of this study. Additionally, the verbal representations of arithmetic facts introduced in Dehaene's Triple Code Model (1992) are proposed to be in the auditory modality – suggesting that the encoding of arithmetic facts during learning might be more linked to auditory representations of number words than their visual counterparts.

All single digit multiplication problems were included except for tie problems (e.g., 2×2) and problems containing 0 and 1 (see Campbell & Graham, 1985), which resulted in a complete set of 56 unique problems. Each problem was repeated four times, twice with a correct solution (e.g., $2 \times 5 = 10$ and $5 \times 2 = 10$) and twice with an incorrect solution (e.g., $2 \times 5 = 15$ and $5 \times 2 = 15$), such that there were a total of 224 trials per operand format (all-digit and auditory-digit), and 448 total trials across the experiment. Incorrect solutions were generated by adding or subtracting 1, 2 or 3 (depending on other constraints) from the first or second operand, making them table-related to the correct answer. Trials were pseudorandomized separately for each operand format. To avoid response bias, consecutive problems never had the same operands or solution, and no more than two correct or incorrect solutions were presented in a row.

For the auditory task, a female fluent speaker of English prerecorded the number words and “times” operation with a standard American accent, using a natural rate of speech and clear diction. Each word was recorded in a padded sound-recording chamber using a DAT recorder. Each number word was spoken in isolation, several times and in different orders, to avoid coarticulation. The clearest example of each word was selected. The sound files were trimmed to the onset and offset of the sound envelope for each individual word, avoiding clipping and distortion, then normalized for amplitude across sound files using Ableton Live 8.

2.1.4 Procedure

2.1.4.1 General: All participants performed both tasks, auditory-digit and all-digit (detailed below), in a single session. Each task consisted of 8 blocks of 28 trials each, with a brief pause between trials, and the order of these blocks was counterbalanced across 3 different lists to control for order effects. The order of the 2 tasks was reversed across participants such that half did the auditory-digit task first. A short practice block that did not include any critical experimental trials preceded each task. Participants were provided a self-timed break between each block and between the tasks to avoid fatigue.

Two sequential operands were presented either as spoken words in the auditory-digit task or Arabic digits in the all-digit task; see Figure 1 for the timing in each task. In both tasks, all solutions were presented as digits with an equals symbol left of center (e.g., “= 10”), such that the numbers always appeared centered on the screen for 350ms. The solutions were followed by a 650ms blank-screen period for ERP collection, then a prompt for a response (“?”) was presented on the screen for at least 350ms and remained until participants indicated whether the provided solution was correct or incorrect. Responses were delayed to avoid motor artifact contamination in the EEG recording. There was a 1 second interstimulus interval (ISI) between the offset of “?” and the next trial, during which participants were permitted to blink.

All visual stimuli were presented in white text on a black background on a 19-inch CRT monitor positioned approximately 1 meter from the participant. All auditory stimuli were presented through ER-1 insert earphones (Etymotic Research, Inc., Elk Grove Village, IL, USA). Participants indicated solution correctness through keyboard responses with the index and middle fingers of the right hand, with finger-response mapping reversed (correct/incorrect) for half of the participants.

2.1.4.2 All-digit trials: The first and second operands were single-digit numbers presented visually in the center of the monitor, each for 450ms with a 250ms blank-screen ISI between them. The second operand was presented with the multiplication symbol to the left of center (e.g., “X 9”). The Arabic numeral solution appeared after a 150ms blank-screen ISI. Each trial started with a 800ms fixation point (asterisk) to fix eye-gaze in the center of the monitor and avoid contaminating the EEG with eye movement.

2.1.4.3 Auditory-digit trials: The first and second operands were presented as individual naturally spoken number words with an average length of 475ms (range: 375 – 639ms). The spoken operation word “times” (690ms duration) was presented between the operands. The Arabic numeral solution appeared in the center of the screen upon termination of the sound file of the second operand. The fixation point (asterisk) was part of the background image, such that in the absence of visual stimulation (e.g., solution and response prompt) it was visible in the center of the monitor to prevent eye movements during auditory stimulation.

2.1.5 Electroencephalogram (EEG) recording—Participants were seated in a dimly lit, sound-attenuating, RF-shielded chamber wearing an elastic cap with 26 tin electrodes embedded in geodesic array (Figure 2). Continuous EEG data was amplified using analog differential bioamplifiers (SA Instrumentation, Co., San Diego, CA) using a bandpass filter

set from 0.01 (RC; -6 dB/oct) to 100 Hz (Butterworth; -12 dB/oct) and sampled at a rate of 250 Hz using the KPCI card on a computer. All electrode impedances were kept below 5 k Ω . The data were referenced online to an electrode placed over the left mastoid process and re-referenced offline to an average of the left and right mastoid electrodes. An additional electrode placed near vertex acted as the ground for all channels. Blinks and horizontal eye-movements were monitored with electrodes placed below and at the outer canthus of each eye (a bipolar recording), respectively.

Epochs of raw EEG data, 100ms prior to stimulus onset until 900ms after stimulus onset, were assessed for inclusion using artifact rejection algorithms with thresholds calibrated by visual inspection for each participant separately. Trials containing blinks, horizontal eye movement, excessive muscle artifact, and EEG drift or blocking were excluded from analysis. Test thresholds were tailored to maximize signal-to-noise ratios per subject. EEG epochs (-100 to 900ms) were then averaged by condition, and a digital bandpass filter (0.1–30.0 Hz, 3rd order Butterworth) was applied prior to analysis (Tanner, Morgan-Short, & Luck, 2015). Data acquisition and data processing was done using ERPSys software (Holcomb lab <https://neurocoglabs.sdsu.edu>). Statistical analysis was performed in R (R Core Team, 2015) using the *ez* package (Lawrence, 2013).

2.2 Results

2.2.1 Behavior—On the offline measure of strategy, participants reported retrieving the answer from memory for approximately 8 out the 10 trials in both the digit operand (0.82) and auditory operand (0.80) conditions (transformation of the problem was the next most commonly reported strategy for both operand types). A correlation of retrieval rates across tasks was very high ($r(24) = 0.84$), indicating that participants reported using similar strategies for both trial types.

Table 1 summarizes participants' offline performance on standardized cognitive measures. We report these standardized measures to facilitate comparison of samples across studies.

During the experimental task, accuracy was high (average 91%; range: 69-100%). There was no difference in performance as a function of operand format or correctness (accuracy in auditory-digit format: correct: 91%, incorrect: 89%, versus the all-digit format: correct: 92%, incorrect: 90%). This was confirmed in an ANOVA with two levels of operand format (digit, auditory) and two levels of correctness (correct, incorrect), which found all p 's > 0.1 (and F 's < 3). In the interest of being inclusive in our sample, no participants were excluded based on performance – lower performance simply resulted in a lower, but sufficient, trial count for ERP analyses.

Delayed response times, measured from the onset of the response prompt, were assessed in an ANOVA with the same factor structure. This found a main effect of correctness ($F(1,25) = 30.06$, $p < 0.001$) and no effect of operand format ($F(1,25) = 0.49$, $p = 0.49$) nor interaction between correctness and operand format ($F(1,25) = 1.41$, $p = 0.25$). The main effect of correctness was due to faster RTs for correct solutions (734.92ms, $SE = 63.25$) compared to incorrect solutions (870.06ms, $SE = 72.51$).

2.2.2 Electrophysiological recordings

2.2.2.1 Waveform morphology: ERPs were measured from the onset of the solution relative to a 100ms prestimulus baseline, and averaged for correct and incorrect solutions, separately, for trials with accurate behavioral responses only. After rejecting trials based on artifacts and accuracy, there was an average of 86 trials (minimum 48, providing adequate signal-to-noise ratio) remaining in each of the critical conditions. This average was the same across tasks and levels of correctness: auditory correct (average: 87; range: 48-109) and incorrect (average: 85; range: 60-107); digit correct (average: 86; range: 51-105) and incorrect (average: 85; range: 54-110).

Figure 3 shows the grand average ERPs for the two tasks. Visual inspection reveals the typical sensory components (P1-N1-P2 over posterior sites and N1-P2 over more frontal sites) in the first 200ms followed by a deviation between correct and incorrect solutions. This deviation appears to peak later for the auditory-digit than the all-digit trials. Importantly, although ERPs are measured relative to each other, and therefore the direction of the voltage deflection is not clear, the effect of congruency appears to be caused by different underlying components across the two tasks. In the all-digit trials, correct solutions elicited a P300-like response (Dickson & Federmeier, 2017), whereas in the auditory-digit trials, the facilitation for correct solutions instead resembles a typical N400 congruency effect, driven by less negativity for correct solutions than incorrect solutions. The all-digit task congruency effect was followed by a second, later effect.

In all the following analyses, Huynh-Feldt corrected p -values and ϵ factors are reported for effects with more than one factor in the numerator. Interactions with electrodes are only discussed when of theoretical relevance.

2.2.2.2 P300/N400 window: To determine the timing of the effect of correctness across the all-digit and auditory-digit formats, we created difference waves (a linear subtraction of the response to incorrect solutions from the response to correct solutions) and measured the maximum amplitude of the resulting differences within a broad window (250-500ms) across all channels in each participant (Figure 4). The effect of congruency peaked about 50ms earlier in the all-digit format (329ms, $SE = 5.5$) than in the auditory-digit format (379ms, $SE = 6.3$). An ANOVA with 2 levels of operand format (digit, auditory) and 26 levels of electrode found this operand format effect to be significant ($F(1,25) = 45.96$, $p < 0.01$).

Average amplitudes were measured using 150ms windows centered over the obtained peak of the effect in each operand format, with a 50ms shift between the time windows used for the all-digit (255-405ms) and auditory-digit (305-455ms) formats (Figure 4). An omnibus ANOVA with 2 levels of operand format (auditory, digit), 2 levels of solution correctness (correct, incorrect), and 26 levels of electrode was performed. This analysis obtained main effects of correctness ($F(1,25) = 88.04$, $p < 0.001$), operand format ($F(1,25) = 25.50$, $p < 0.001$), and an interaction between correctness and operand format ($F(1,25) = 9.71$; $p < 0.01$).

The effect of correctness was due to a more positive response to correct solutions ($\mu V = 2.27$, $SE = 0.26$) than incorrect solutions ($\mu V = 0.87$, $SE = 0.24$), and the effect of operand format was due to more positive responses to digit solutions when preceded by digits (all-

digit format, $\mu V = 2.64$, $SE = 0.36$) than when preceded by auditory operands (auditory-digit format, $\mu V = 0.50$, $SE = 0.26$). The interaction between correctness and operand format was due to there being a larger effect of correctness in the all-digit format (correct: $\mu V = 3.58$, $SE = 0.38$; incorrect: $\mu V = 1.69$, $SE = 0.37$; within-subject difference $\mu V = 1.90$, $SE = 0.20$) than in the auditory-digit format (correct: $\mu V = 0.95$, $SE = 0.28$; incorrect: $\mu V = 0.05$, $SE = 0.30$; within-subject difference $\mu V = 0.90$, $SE = 0.24$).

The omnibus ANOVA also revealed interactions between correctness and electrode ($F(25,625) = 9.03$, $p < 0.001$, $\epsilon = 0.13$), operand format and electrode ($F(25,265) = 4.75$, $p < 0.001$, $\epsilon = 0.19$), but, critically, no interaction between correctness, operand format, and electrode ($F(25,625) = 0.99$, $p = 0.41$, $\epsilon = 0.15$). An additional analysis including distributional factors of laterality (lateral, medial), hemisphere (right, left), and anteriority (prefrontal, frontal, parietal, occipital), revealed significant interactions of correctness with laterality ($F(1,25) = 29.52$, $p < 0.001$), anteriority ($F(3,75) = 6.26$, $p < 0.05$, $\epsilon = 0.43$), and both laterality and anteriority ($F(3,75) = 4.91$, $p < 0.01$, $\epsilon = 0.90$). The interactions reflect the typical effect of arithmetic correctness (using mastoid references) with largest amplitude at centromedial and parietal electrodes (see Figure 3 for topographic plots of these effects).

2.2.2.3 Effects of correctness between 400-700ms—Figure 3 shows a second modulation based on correctness primarily for the all-digit condition immediately following the N400/P300 window. Mean amplitude was measured for each condition in equivalent 200ms windows immediately following the earlier P300/N400 measurement window: 405-605 for all-digit (centered around the peak of the effect, ~500ms based on peak latency measurement) and 455-655 for auditory-digit. These data were subjected to an omnibus ANOVA with 2 levels of correctness, 2 levels of operand format, and 26 electrodes. The ANOVA identified an interaction between correctness and operand format ($F(1,25) = 8.35$, $p < 0.01$), as well as a marginal main effect of correctness ($F(1,25) = 4.12$, $p = 0.053$).

The interaction between correctness and operand format was explained by a large effect of correctness in the all-digit format (correct: $\mu V = 0.26$, $SE = 0.25$; incorrect: $\mu V = 1.17$, $SE = 0.30$; within-subject difference: $\mu V = -0.90$, $SE = 0.28$; $F(1,25) = 10.43$, $p < 0.01$) and no significant effect in the auditory-digit format (correct: $\mu V = 0.89$, $SE = 0.26$; incorrect: $\mu V = 0.94$, $SE = 0.35$; within-subject difference: $\mu V = -0.04$, $SE = 0.27$; $F(1,25) = 0.02$, $p = 0.88$). The mean amplitudes reveal that, unlike during the earlier P300/N400 time window, the effect of correctness in the all-digit condition here reflected a less positive/more negative response to correct solutions than to incorrect solutions.

2.3 Interim Discussion: Experiment 1

We observed morphological differences in the effect of correctness elicited by the solution to multiplication problems presented with digit operands or spoken operands. Digit operands elicited a larger, earlier, and more positive-going effect than spoken operands did. Critically, these differences were due to the operand format, given that the solution format was held constant. Thus, differences at the solution reflected differences in the preparatory process for retrieving and verifying the math facts.

Visual inspection of the wave morphology for the correct and incorrect trials in each condition revealed a positive-going deflection for the correct trials for all-digit condition and a negative-going deflection for the incorrect trials for digits preceded by spoken operands. Indeed, the obtained 50ms latency shift between the two conditions is consistent with the timing of a P300 and N400, respectively. Note that this difference was not due to the switch between formats in the auditory-digit condition because very similar effects have been reported for problems presented as words (Martinez-Lincoln et al., 2015; Salillas & Wicha, 2012). In these studies, both the solution and the operands were presented as written words. Indeed, the fact that the effect at the solution was strikingly similar for digits preceded by words in the current study and for written words preceded by words in earlier studies suggests that the presence of the N400 effect is driven by the format of the preceding context (operands) rather than the format of the solution.

A subsequent effect between 400-600ms after solution onset was observed only when the solution was preceded by digit operands, with correct responses eliciting more negativity than incorrect responses. A similar effect had been previously reported as a late positive component/complex (LPC) by some (Dickson & Federmeier, 2017; Domahs et al., 2007) and as a P300 by others (e.g., Avancini et al., 2015, Jost et al., 2004; Niedeggen & Rösler, 1999; Núñez-Peña & Escera, 2007) for incorrect solutions. The incorrect solutions in the present study were relatively difficult to categorize as incorrect because they were divisible by one of the preceding operands (a property commonly referred to as “table related”) and were numerically close to the correct solution. Thus, it is unlikely that this effect was a positivity driven by the salience or obvious implausibility of the incorrect solutions (Dickson & Federmeier, 2017; Jost et al., 2004; Núñez-Peña & Escera, 2007). Indeed, in the present work the late effect appears instead to be the result of a slow wave negativity for the correct solutions that often follows the initial P300 (Duncan-Johnson & Donchin, 1977).

In contrast to solutions preceded by digit operands, the morphology of the waveforms and the timing and distribution of the correctness effect for solutions preceded by spoken operands is highly consistent with an N400 component, with larger amplitude for incorrect than correct trials. That both the P300 and the late effect were jointly absent here supports the interpretation that those effects (for correct solutions preceded by digits) are indeed related to one another. On the basis of the present within-subject evidence and the broader ERP arithmetic literature, it seems fair to propose that the use of all digits promotes an early P300-like response to correct answers (and a later reversal with a more positive response to incorrect solutions), whereas use of spoken number words for operands promotes more traditional language-like N400 effects.

However, there is one particularly strong alternative explanation for these results: perhaps participants simply needed more time to prepare for the incoming visual solution when the mathematical context was presented in the auditory modality. It is possible that giving listeners more time would allow them to view the digit solution preceded by spoken words as a target item, resulting in a P300 for correct answers. Reaction times did not significantly differ as a function of operand modality, which might argue against this possibility, but participants had to delay their responses until a prompt. Early processing differences might thus have been neutralized during the lag before the prompt. In Experiment 2, we

deliberately manipulate the length of time between the presentation of the second operand (e.g., the complete problem) and the presentation of the solution, with both a shorter delay (matched in time to what was used for numeral contexts in Experiment 1) and a much longer delay (1 second) to provide participants with excessive time to access/retrieve the correct solution (see Figure 5 for the timing structure).

3. EXPERIMENT 2

3.1 Method

3.1.1 Participants—Twenty-six new participants from the same local University population as Experiment 1 were included in analyses (2 additional subjects were excluded due to technical errors). All participants were right-handed (determined by Edinburgh handedness inventory; average: 0.93; range: 0.7-1; Oldfield, 1971), native monolingual English speakers (10 male; mean age: 20.2; range: 18-24) with normal or corrected-to-normal vision and none had cognitive or neurological impairments (including history of brain injury) or current use of psychoactive medications. Participants self-reported equivalent education levels, exposure to math, and exposure to language as the participants included in Experiment 1. Informed consent was received in accordance with the IRB at the University of Texas at San Antonio.

3.1.2 Offline behavioral cognitive assessments—In Experiment 2, measures of phonological awareness and working memory were obtained using the same standardized assessments as Experiment 1 (see Table 1). Math fluency was measured using an equivalent test to Experiment 1, a subtest of the Wechsler Individual Achievement Test in which participants solve single-digit multiplication problems within a 60-second time limit (Math Fluency-Multiplication, WIAT-III, Wechsler, 1992).

3.1.3 Stimuli and procedure—The same physical stimuli were used from the auditory-digit task in Experiment 1 with three changes: the fixation point was changed from an asterisk to a multiplication “X”, all number words were normalized to 450ms in duration, and the operand word “times” was removed so that the operands were spoken sequentially with no intervening word stimulus. These changes decreased the total duration of each trial and made the auditory and prior visual stimuli a closer match in timing. Other than these revisions, the same trial structure was adopted from Experiment 1 (consult Figure 1 versus Figure 5 to compare). The critical manipulation in Experiment 2 was the duration of the ISI between the second operand and the solution. In the short ISI condition, the ISI was 150ms, mimicking the timing of the all-digit condition from Experiment 1 as closely as possible. In the long ISI condition, there was a 1000ms delay between the offset of the second operand and the onset of the digit solution.

3.1.4 Electroencephalogram (EEG) recording—Recording parameters were identical to Experiment 1, with the exception of using a different bioamplifier and electrodes. Continuous EEG was recorded using a BioSemi bioamplifier running ActiveView software and 26 active electrodes containing an Ag-AgCl sintered electrode pellet fitted in a custom electrode cap (Electro-Cap International Inc. using BioSemi electrode holders). Electrodes

were geodesically arranged, as in Experiment 1 (Figure 2). A signal from each electrode was recorded with respect to a Common Mode Sense (CMS/DRL) active/ground electrode, and was referenced offline to the average of two electrodes placed over the right and left mastoid processes. Buffer amplifiers contained within the electrodes reduce any noise induced by the amplification and digitization system (BioSemi ActiveTwo, BioSemi B.V., Amsterdam) allowing for measurements of body surface potentials with high electrode impedances. Electrode offsets were kept low (below 50 millivolts). A fixed first order analog anti-aliasing filter with a half-power cutoff 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 $\frac{1}{8}$) using a 5th order sinc digital filter to avoid aliasing.

ERPLab software³ was used to process and measure the data (Lopez-Calderon & Luck, 2014). Epochs of raw EEG data (–100 to 900ms) were assessed for inclusion using artifact rejection algorithms with thresholds calibrated by visual inspection for each participant separately. These algorithms were then applied to the entire data set to exclude trials containing blinks, horizontal eye movement, excessive muscle artifact, and EEG drift or blocking. EEG epochs were then averaged by condition, and 2nd order Butterworth digital filters with a low cutoff at 0.1 Hz and a high cutoff at 30 Hz were applied before analysis.

3.2 Results

3.2.1 Behavior—Accuracy on the ERP tasks was high (average 92%; range: 60-100%). There was no difference in performance as a function of ISI or correctness (accuracy in short ISI, correct: 92%, incorrect: 91%; long ISI: correct: 91%, incorrect: 92%). This was confirmed in an ANOVA with two levels of ISI (long, short) and two levels of correctness (correct, incorrect), which found all p 's > 0.1 (and F 's < 2). Behavior on offline standardized tests is reported in Table 1 for comparison with the participants in Experiment 1. No participants were excluded as a result of performance.

Delayed response times across critical conditions were also measured as in Experiment 1. This found a main effect of correctness ($F(1,25) = 52.07$, $p < 0.001$) and no effect of ISI ($F(1,25) = 3.82$, $p = 0.06$) nor interaction between correctness and ISI ($F(1,25) = 1.51$, $p = 0.23$). The main effect of correctness was due to faster RTs for correct (408.09ms, $SE = 19.63$) compared to the incorrect (477.11ms, $SE = 22.54$) solutions.

3.2.2 Electrophysiological recordings

3.2.2.1 Waveform morphology: As in Experiment 1, ERPs were measured from the onset of the solution relative to a 100ms prestimulus baseline, and averaged for correct and incorrect solutions, separately, for trials with accurate behavioral responses only. After artifact rejection, an average of 93 items (minimum: 48) were included in each of the critical conditions (accurately identified correct and incorrect solutions in each ISI condition). This average was the same across ISI and levels of correctness: short ISI correct (average: 93;

³The use of different software packages in Experiment 1 and 2 was necessitated by the bioamplifier employed. This is not a concern for the results as ERPLab and ERPSystem are very similar software packages, both originating from the Steve Hillyard Lab software (ERPSS); critical differences are noted. Notably, the lack of discernible differences in the results across experiments for the auditory-digit conditions demonstrates that the two systems used are highly compatible.

range: 52-111) and incorrect (average: 92; range: 62-111); long ISI correct (average: 94; range: 49-112) and incorrect (average: 93; range: 49-112).

Figure 6 shows the grand average ERPs. Visual inspection reveals the typical sensory components P1-N1-P2 over posterior electrodes and N1-P2 over anterior sites in the first 200ms. The sensory components are followed by an effect similar to what was seen in the auditory-digit task of Experiment 1, with an increased negativity to incorrect solutions compared to correct solutions, characteristic of an N400 effect. This effect appears to occur earlier for the long than short ISI conditions.

In all of the following analyses, Huynh-Feldt corrected p -values and ϵ factors are reported for effects with more than one factor in the numerator. Interactions with electrode are only discussed when of theoretical relevance.

3.2.2.2 N400 window: In order to determine the timing of the effect of correctness across the short and long ISIs, we created difference waves (as in Experiment 1) and measured the maximum amplitude of the resulting differences within a broad window (250-500ms) across all channels in each participant. The effect of correctness peaked about 40ms earlier in the long ISI format (376ms, $SE = 6.3$) than in the short ISI format (415ms, $SE = 6.7$). An ANOVA with within-subject variables of 2 levels of ISI length (short, long), and 26 levels of electrode revealed a significant effect of ISI length ($F(1,25) = 15.72$, $p < 0.01$).

Average amplitudes were measured in 200ms windows centered over the peak difference obtained for each ISI condition in the latency analysis. This resulted in 2 time windows shifted by 40ms: short ISI (315-515ms) and long ISI (275-475ms).

An initial omnibus ANOVA with 2 levels of ISI, 2 levels of correctness (correct, incorrect), and 26 levels of electrode was performed. This obtained main effects of correctness ($F(1,25) = 32.74$, $p < 0.001$) and ISI ($F(1,25) = 20.85$, $p < 0.001$). The effect of correctness was due to larger negative amplitude for incorrect solutions ($\mu V = 2.37$, $SE = 0.39$) than correct solutions ($\mu V = 3.52$, $SE = 0.51$). An interaction between correctness and electrode ($F(25, 625) = 7.92$; $p < 0.001$, $\epsilon = 0.21$) was examined using a separate ANOVA with the same distributional factors used in Experiment 1 (2 levels of hemisphere, 2 levels of laterality, and 4 levels of anteriority). This distributional analysis found that the effect of correctness interacted with laterality ($F(1,25) = 19.95$, $p < 0.001$), anteriority ($F(3,75) = 5.50$, $p < 0.05$, $\epsilon = 0.49$), and laterality and anteriority ($F(3,75) = 4.49$, $p < 0.001$, $\epsilon = 0.90$). The effect of correctness reflected the typical distribution for an N400 congruency effect, with its largest size over centromedial electrode sites (see Figure 6).

The main effect of ISI was due to less negative amplitude overall for the long ISI ($\mu V = 3.89$, $SE = 0.49$) than for the short ISI ($\mu V = 2.00$, $SE = 0.48$). Within-subject effect contrasts (i.e., difference waves) found similar-sized effects of correctness in the short ISI format ($\mu V = 1.04$, $SE = 0.24$) and the long ISI format ($\mu V = 1.25$, $SE = 0.26$). The interaction between correctness and ISI did not reach significance ($F(1,25) = 0.49$, $p = 0.49$), indicating that the effect of correctness on the amplitude of the N400 was equivalent across short and long delays.

3.2.2.3 Sustained effect: Unlike Experiment 1, there was a sustained increase in negative amplitude for incorrect relative to correct solutions. An omnibus ANOVA with 2 levels of correctness, 2 levels of ISI and 26 levels of electrode was conducted on mean amplitudes of the last 300ms of the extracted time window (between 600-900ms). The analysis confirmed a significant main effect of correctness ($F(1,25) = 10.25, p < 0.01$), with a larger negative amplitude to incorrect ($\mu V = 1.49, SE = 0.46$) than correct ($\mu V = 2.20, SE = 0.55$) solutions (within-subject difference $\mu V = 0.71, SE = 0.41$). There was no effect of ISI, as a main effect ($F(1,25) = 2.71, p = 0.11$) nor in interaction with correctness ($F(1,25) = 0.67, p = 0.42$), in this window.

3.3 Interim Discussion: Experiment 2

In Experiment 1, ERPs measured at the solution following spoken operands showed that participants distinguished between correct and incorrect solutions at the level of semantics, as reflected by a typical N400 congruency effect. It is plausible that participants were reading these arithmetic problems as if they were reading meaningful sentences, where expected continuations elicited less negativity on the N400 than unexpected words (indeed, this is the original interpretation of ERP correctness effects). However, it was not obvious how best to interpret the lack of a distinctive P300 for correct solutions. Performance in accuracy and response times showed no indication that participants' comprehension of the multiplication problems was impaired compared to comprehension of digit operand trials. One possibility was that the temporal unfolding of the spoken number words allowed less time (than visually presented digits) to anticipate the solution as the target. Target P300s are highly sensitive to internally generated factors like subjective probability, confidence, and level of attention (Polich, 2007, 2012), and lacking sufficient time to generate visual representations of the expected solution might have prevented participants from recognizing and interpreting the provided correct digit solutions as targets.

In Experiment 2, to address this possibility, we deliberately manipulated the amount of time between the second operand and the solution. Again, there was no evidence of a P300 in response to correct solutions – even when participants had a full second between the end of the sound for the second operand and the onset of the visual solution. At the longer delay, there was a morphological hint of a positive component around 330ms at medial parietal recording sites (see Figure 6), but there was no distributional shift or statistically significant increase in the size of the effect that would be consistent with significant new contributions from a P300. There also was no LPC/P300 for incorrect solutions, which have been observed following the P300 for correct solutions (e.g., Domahs et al., 2007; Niedeggen & Rösler, 1999).

Instead, as in Experiment 1, the digit solutions preceded by spoken operands elicited only an N400 component with amplitude reflecting ease of access to the meaningfulness of the stimulus (i.e., long-term semantic memory representations). Unexpected, incorrect solutions were inconsistent with the semantic memory representations that had been activated by the preceding context, thus demanding more activation of new semantic information and resulting in a larger (more negative) N400 response than correct (expected) solutions. Experiment 2 demonstrated that the alternation in the underlying componentry of ERP

correctness effects between an N400 to a P300 was not simply about amount of time to generate expectations. Rather, the interpretation of the task at hand appears to be influenced by the format of the operands, which we address with more detail in the General Discussion.

Delaying the solution presentation did have one effect: the peak of the N400 effect appeared later when there was less time between the operand and solution. This latency difference is distinct from the latency difference in Experiment 1, where the effect of correctness in all-digit trials was too early to reflect access to semantic processing (see Federmeier, Kutas & Dickson, 2015, for an extended review of the timescale of levels of information processing across stimulus types of differing complexities and modalities). The N400 is quite stable in time across modalities, and it is rare to find significant shifts (Kutas & Federmeier, 2011). Within-subject delays have been obtained under conditions where the system is taxed (e.g., due to very rapid sentence presentation of 10 words/second, Kutas, 1987; cf. Wlotko & Federmeier, 2015). More often, N400 latency differences are reported between subjects in clinical populations where processing speed might be impaired (e.g., in low-fit children, Scudder et al., 2014; in older adults, Wlotko, Lee, & Federmeier, 2010) or in healthy populations with individual differences in their semantic networks (e.g., vocabulary size, Moreno & Kutas, 2005). Here we report a delay of N400 latency with a simple within-subject manipulation of giving less time to prepare for an arithmetic solution. An alternate account of the N400 shift here is that it might be reflecting earlier latency of a slow and variable P300 for correct solutions in the long ISI condition that is not particularly visible in the plots. This is possible, but it is challenging to reconcile this account with the behavioral data, which did not distinguish between ISIs. Thus, although manipulating ISI does not appear sufficient to radically alter the type of processing participants engage in, particularly fast presentation of items does slow the speed with which they are able to process the semantic content of the solutions (cf., Domahs et al., 2007, who found similar results with a stimulus onset manipulation in a paradigm using all Arabic numerals).

Although there was again no evidence for the later reversal of the correctness effect observed for all-digit trials (where correct solutions become more negative than incorrect solutions), the conditions did show a difference in the form of a sustained negativity to incorrect responses. This negativity started after the N400 time window and remained through the recording epoch. A possible interpretation of this difference in amplitude driven by correctness might be that it reflects differential engagement of working memory resources across answer types, as sustained negativities have also been seen in response to ambiguity resolution and joke comprehension (although these effects are also typically more frontally distributed) (Kluender & Kutas, 1993; Lee & Federmeier, 2009; Ruchkin et al., 1992; Wlotko & Federmeier, 2012). Another explanation for this effect is that it is a continuation of the N400. As in the language and memory literature, additional work is necessarily to clarify the nature of late effects in this domain.

4. General Discussion

The goal of this study was to determine if number format affects the brain's response to solving multiplication problems, and, by implication, the underlying cognitive processes involved. Namely, can the format of the multiplication operands, spoken number words or

Arabic digits, affect the way the brain assesses identical digit solutions? In Experiment 1, we measured ERPs to the onset of Arabic numeral solutions preceded by two sequential operands presented as either Arabic digits (all-digit condition) or spoken number words (auditory-digit condition). Robust effects of correctness were observed in both cases, with correct solutions eliciting more positive amplitude and incorrect solutions eliciting more negative amplitude. However, the underlying waveform morphology and timing of the effects were clearly distinct. Digit operands elicited a P300 component for correct solutions peaking at 329ms, consistent with previously reported target P300 effects (Dickson & Federmeier, 2017; Jasinski & Coch, 2012), while spoken number words elicited a larger N400 component for incorrect solutions peaking at 380ms, in line with previously reported semantic N400 effects (Kutas & Federmeier, 2011; Martinez-Lincoln et al., 2015; Salillas & Wicha, 2012).

In Experiment 2, we tested if the lack of a P300 might have been due to participants needing more time to process the problems when presented as spoken words. We manipulated the duration of the ISI between the second spoken operand and the digit solution to be either 150ms, which was the same as the all-digit ISI in Experiment 1, or 1000ms. There was no discernible difference in the underlying waveform morphology between the ISI conditions, indicating that even with a full second delay the brain still did not treat the digit solutions as “targets” when preceded by spoken words. Moreover, the size of the effect was equivalent for both ISIs, which would not be expected if there had been additional contributions from a P300 in the longer ISI. There was, however, a timing difference, with a later N400 effect when duration between the second operand and solution was very short, similar to what has been reported in cases where efficiency of semantic access appears to be compromised (Scudder et al., 2014; Wlotko et al., 2010).

Of importance across the two experiments is that this N400 effect was observed regardless of the latency between the second operand and the solution, from no delay (Experiment 1) to a full second delay (Experiment 2) (see Figure 7). This indicates that the N400 modulation is in line with what has been observed in the sentence processing literature, where N400 effects reflect automatic semantic processing elicited by contextual expectations (Connolly, Stewart, & Phillips, 1990; Deacon & Shelley-Tremblay, 2000; DeLong, Urbach, & Kutas, 2005; Federmeier & Kutas, 1999a, 1999b; cf., Wicha, Moreno, & Kutas, 2004; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). Minimal time between the second operand and proposed solution was still sufficient for the brain to differentiate correct from incorrect solutions at relatively early stages of cognitive processing (i.e., starting around 200-300ms poststimulus), even when the modality of the operand and solution differed. This is not particularly surprising given that college students have years of experience retrieving answers to these simple multiplication problems (albeit sometimes with external aids). Critically, however, this N400 effect was absent with digit operands. Before discussing the significance of these findings, we will first unpack the specific evidence we found in support of interpreting the componentry in this manner.

4.1 Component Identification

In Experiment 1, the effect of correctness for solutions preceded by digit operands was significantly earlier and larger than the effect of correctness for the same solutions preceded by operands in the auditory modality. The timing (peaking around 330ms) and morphology (a distinct positivity for correct items, see left plots in Figure 3) of the correctness effect following digit operands is consistent with a P300 effect (Polich, 2012). In contrast, the timing (around 400ms) and morphology (negative-going waveforms) of the auditory-digit correctness effect are consistent with typical N400 sentential congruency effects (Kutas & Federmeier, 2000).

Interestingly, the response to solutions with digit operands also featured a later effect, where the conditions reversed (more negative for correct solutions). In the language literature, the precise significance of effects following the N400 are a matter of some debate (see Van Petten & Luka, 2012, for review), and in the context of N400 effects, this would likely be interpreted as an LPC for the incorrect (incongruent) solutions. In the arithmetic literature, later positivities on incorrect solutions have also been interpreted as LPCs or delayed P300s (Dickson & Federmeier, 2017; Domahs et al., 2007; Jost et al., 2004; Niedeggen & Rösler, 1999; Núñez-Peña & Escera, 2007). Here, since slow-wave *negativities* commonly follow P300s (Duncan-Johnson & Donchin, 1977), a simple account is that correct answers elicit a robust P300 complex that also includes a later negativity. Neither the target P300 nor the subsequent negativity is present in our work using auditory operands (indeed, Experiment 2 found a later, sustained negativity for incorrect solutions instead).

These outcomes collectively support our claim that a fundamentally different type of cognitive processing was undertaken by participants when viewing and judging solutions preceded by digit operands compared to when they judged the same solutions preceded by auditory operands. We will begin by discussing the significance of the results in the digit operand task and then discuss the N400 effect in the paradigms utilizing auditory operands.

4.2 Response to Digit Solutions Preceded by Digit Operands: Target P300

A P300 was observed in response to correct digit solutions when preceded by digit operands relative to the incorrect solutions. We attribute this brain response to participants perceiving the correct solutions as target items in a binary item categorization task. The arithmetic verification task itself resembled a typical eliciting paradigm for P300s because the critical solutions were items on which participants rendered simple categorical judgments (Polich, 2007, 2012). Further, the task itself centered on recognition of correct solutions, which were preexperimentally familiar to participants. Thus, correct solutions were more salient and easily categorized than incorrect ones, eliciting a larger P300. Consistent with this view, reaction times were faster for correct solution verification than for incorrect solution rejection. We are not the first to characterize this positivity as a P300 (Jasinski & Coch, 2012). The target P300 here was likely facilitated by participants' familiarity with the physical format used in the stimulus presentation and their years of exposure to the experimental items. Sentence-reading paradigms with plausibility judgments on the final word similarly elicit positivities for expected/congruent words, in addition to typical N400 congruency effects (e.g., Geyer, Holcomb, Kuperberg, & Perlmutter, 2006; Kolk, Chwilla,

Van Herten, & Oor, 2003). In paradigms utilizing written number words as solutions to arithmetic problems, there appear to be positivities for correct answers, although these studies focused on the co-existing N400 effect and not the positivity (e.g., Fisher et al., 2010). It is important to note that P300s are not an obligatory response to a correct or expected completion: sentence-processing paradigms that lack an imposed plausibility judgment on the final word elicit an N400 effect and no target P300 (e.g., Federmeier & Kutas, 1999a,b; Federmeier et al., 2002).

Unlike the P300, the N400 is an obligatory response for any potentially meaningful item, in any perceptual modality. However, although participants do generate an N400 *component* in response to Arabic numerals, in the current study we did not obtain any direct evidence of an *effect* on this component itself when the preceding operands were digits due to its early timing (as in Dickson & Federmeier, 2017). The absence of the N400 effect may be most easily accounted for by comparisons to results obtained in other domains looking at discontinuities in patterns (e.g., musical melodies with an unexpected note). When familiar melodies are presented, there is no modulation of the N400 as a function of expectancy (Paller et al., 1992). This suggests that these learned sequences do not automatically contact and modify distributed, meaningful semantic representations. Similarly, single-digit Arabic numerals, absent other cues linking to broader semantics, might not succeed in contacting and modifying distributed representations of meaning either (cf., Galfano et al., 2004, 2009). Thus, there is no substantial facilitation of the N400 for expected items and instead the information value conveyed by the answer types is primarily reflected on modulations of the P300 complex.

A more complex interpretation of the absence of the N400 effect and the presence of the P300 effect can be found by looking at the second language learning literature. Early in second language learning, grammatical errors are reflected on the N400 instead of the P600 (a positivity typically elicited by syntactic errors), suggesting that participants comprehend that a word was unexpected but are unable to attribute the mistake to the level of grammar (McLaughlin et al., 2010; Tanner, McLaughlin, Herschensohn, & Osterhout, 2013). However, as the same participants gain increasing familiarity with the second language, these errors begin to elicit positivities (P600s) instead (a complicated process overviewed by Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, and Molinaro, 2006). This switch is thought to reflect participants' growing facility with their newly acquired language's grammatical structure. The errors (incorrect solutions) in the arithmetic task are not syntactic or grammatical in nature. The P600 itself is part of a broader debate regarding the significance of a *family* of positivities – the P300, LPC, and P600 – which are collectively associated with explicit recognition and/or deliberative processing of stimuli compared to the more *implicit* comprehension of stimuli that can elicit N400 effects (for review of positivities, see Van Petten & Luka, 2012). Under this explanation, in earlier stages of arithmetic comprehension, correctness effects would be reflected on the N400. Only after extensive familiarity with the expressions would the processing of answer type be instead reflected on ERP components more linked to explicit recognition processes (i.e., P300).

4.3 Response to Digit Solutions Preceded by Auditory Operands: N400 Effects

Our finding that solution correctness is processed at the level of meaning (N400) when the operands are in auditory format is notable because participants theoretically could have responded with a P300 complex just as when the operands were digits. Instead, even though the task was the same, the distinctive positivity for correct answers was absent. Use of the less-familiar auditory operands may have shifted the subjective interpretation of the task from one of target detection to one of basic comprehension. Although participants were actively judging the answers for correctness, this reinterpretation of the task to one of comprehension may have been further facilitated by requiring delayed responses, perhaps lowering the urgency of the task (Polich, 2012). Just as P300s can be elicited in sentence comprehension studies by inclusion of an explicit plausibility judgment on the final word, it is possible that emphasizing speed (urgency) of responses might similarly result in more distinct P300s here. In the present results, however, the brain responses in this task resembled the N400 effect that is observed in passive sentence listening experiments (Federmeier et al., 2002; Holcomb & Neville, 1991). Consistent with the present result, N400 effects in language studies are robust to format switching. For example, written and spoken sentences that terminate in line drawings similarly obtain modulations on the N400 as a function of conceptual expectancy (Wicha, Bates, Moreno & Kutas, 2003; Wicha, Moreno & Kutas, 2003; Federmeier & Kutas, 2001).

The modulation on the N400 is further evidence that contextual support can facilitate interpretation of a given item as meaningful when, in other experimental contexts, it was not perceived as meaningful (i.e., would not elicit N400 effects). For example, we previously mentioned that unexpected notes in melodies or color patches do not elicit N400 effects. However, with more enriched semantic contexts (i.e., preceding linguistic stimuli), N400 effects can be obtained in response to these types of stimuli (Brang et al., 2008; Daltrozzo & Schön, 2009).

An alternative explanation is that format-switching itself might reduce the inclination to interpret visual items as targets of the auditory problem. Under this interpretation, presenting operands in the auditory format may have caused participants to be unprepared for immediate, confident recognition of solutions in their visual numeral format. Within the arithmetic ERP processing literature, there are no other studies involving format switching from which to reference with respect to the presence/absence of the P300 (cf., Szűcs & Csépe, 2004, who measured other ERP components in a format-switching study). However, the P300 tends to be more robust than this interpretation affords, with difficult categorization tasks resulting in delayed and/or smaller P300s rather than their absence entirely (Polich, 2012). Moreover, in studies that use all written number word stimuli, with no format switching across operand and solution, the (unreported) positivity for correct trials has typically been much less salient than the N400 effect (e.g., Martinez-Lincoln et al., 2015; Salillas & Wicha, 2012).

In sum, it is possible that another effect of using auditory number words, other than eliminating the P300, was stimulating additional activations of semantic memory representations that led to sensitivity on the N400 as a function of correctness. That is, perhaps the context provided by Arabic numeral operands facilitates detecting the correct

solution as the target, whereas utilizing words facilitates the construction of expectations at the level of semantics. Whether these findings hold under other manipulations of the task or stimuli, such as requiring urgent responses or changing the solution format independently of operand format, will need to be further investigated.

4.4 Implications for the Triple Code Model of Math Cognition

Under the most dominant model of math cognition, the Triple Code Model (TCM; Dehaene, 1992), numbers can be represented by three separate codes: an amodal representation of magnitude (i.e., numerosity), a verbal code of number words, and a visual code of Arabic numerals. This framework has been extraordinarily influential in motivating different levels of inquiry regarding numerical and mathematical representations of knowledge (i.e., with 1853 citations on Google Scholar). In the interest of speaking to the broader math cognition literature, we relate our findings to the predictions made by this model.

In the TCM, overlearned arithmetic facts, like the simple multiplication problems used in this experiment, are stored and accessed as rote verbal memories within the verbal code. Thus, arithmetic problems presented as Arabic numerals necessarily contact the verbal code to retrieve correct answers from memory. If the verbal code is not accessed, a calculation would need to take place to generate the answer because there is not an encoded form of the arithmetic problems in any other source. By implication, arithmetic facts presented in the written, auditory, and digit formats should all similarly engage cognitive processes to retrieve the facts from a shared verbal memory system.

Our results challenge this aspect of the proposal. Participants did report utilizing the same strategy across operand formats in an offline assessment, claiming to retrieve (“remember”) answers ~80% of the time, and their behavior on the math task was similar across formats. From both their actual and self-reported behavior, it might be reasonably inferred that both formats lead to similarly effective fact retrieval from verbal memory. However, the ERPs revealed critical differences as a function of operand format, suggesting instead that different cognitive processes underlie these similar behavioral reports and outcomes.

An N400 effect was observed for solutions preceded by spoken words. Such effects most famously emerge in language contexts, so what we obtained might appear like direct evidence of the specific auditory “verbal” memory access proposed in the TCM. Importantly, the N400 component is sensitive to meaningfulness across all sensory modalities (Kutas & Federmeier, 2011). Therefore, although this effect is consistent with verbal memory access as proposed in the TCM, it is also consistent with access and manipulation of semantic memories in other (e.g., written) forms as well. In contrast, when operands were in the digit format, there was no clear evidence of an independent N400 effect of correctness, so semantic memory does not appear to have been critically involved in the answer retrieval process. The obtained P300 effect, however, necessarily indicates that participants accessed memories for these arithmetic facts and built expectations for correct solutions *at some level*, perhaps just not semantic. It is possible that, when arithmetic facts were accessed during presentation of auditory number word operands, semantic-level expectations were constructed and reflected as an N400 effect on the solution, but access to arithmetic memories involved more superficial levels of processing when familiar digits

were shown. Participants might thus perceive that they retrieved the solutions directly from memory in all cases, even though the level (or type) of processing involved differed across operand formats.

Another relevant point is the timing of the correctness effects across operand formats, with the P300 effect occurring earlier than the N400. In the behavioral literature, answer verification is earlier (faster RTs) to problems presented as Arabic numerals compared to verification times for arithmetic problems using written words (Frenck-Mestre & Vaid, 1993). Timing differences of this nature have been a historical point of contention with the TCM, which appears to suggest that access to memorized arithmetic-related information should instead be faster for language representations than Arabic numeral representations. Our data may be more in line with models of math cognition that incorporate experience-dependent development, for both speed and type of brain response, into the strength and nature of arithmetic representations (e.g., Campbell & Epp, 2004; Szűcs & Csépe, 2004).

4.5 Significance of Component Identification for the Domain of Math Cognition

It should be evident that identifying whether the effect of correctness is primarily driven by a P300 or an N400 effect is important for reasons beyond ERP naming conventions. The P300 is a fundamentally different neurological signal than the N400, with a very different psychological significance. Theories about the P300⁴ include suggestions that it reflects updating of context in working memory (an enduring theory put forth by Donchin and Coles (1988)), that it reflects stimulus categorization, or that it reflects decision monitoring, often making similar predictions about its timing and size (Polich, 2012). The P300 also has a different proposed neurobiological source. Specifically, the P300 has been associated with activity in the norepinephrine-locus coeruleus system that projects to temporal-parietal areas (Nieuwenhuis, Aston-Jones, & Cohen, 2005). The N400 instead is directly associated with semantic memory (i.e., stimulus meaningfulness) and its sources are thought to be more broadly distributed, although perhaps especially coming from areas of the temporal lobe (Kutas & Federmeier, 2011).

Because the P300 and N400 are both well-characterized and relatively well-understood phenomena, attributing an obtained correctness effect to one or the other critically determines not only how one interprets their present finding, but also how one goes forward in designing studies to directly manipulate the size and timing of the effect in future studies. For example, under the assumption that the correctness effect is driven by a large target P300 for correct solutions (in experiments utilizing all Arabic numerals), one can then take advantage of what is known about the P300 to ask appropriate questions targeting the sensitivities of the measure. The size and/or latency of the P300 are widely used as biomarkers of cognitive ability in clinical contexts, and can be influenced by factors varying from arousal and attention to corpus callosum size and handedness (Polich & Criado, 2006). Reports of ERP correctness effects in arithmetic have not included reports of correlations with behavioral measures (e.g., Niedeggen et al., 1999; Domahs et al., 2007).⁵ Perhaps there is a misalignment between the response properties of the derived ERP correctness effect and

⁴In this report, our references to the P300 are to the P300b, and not the P300a or “novelty P300”. That component has different eliciting conditions, psychological relevance, and neurobiological sources than the P300b (Polich, 2007, 2012).

the cognitive processes typically captured by standardized behavioral tests (e.g., arithmetic subtests of Woodcock-Johnson). For example, participants' enthusiasm for math and for the experiment at the time of test might be a more reliable predictor of the size of P300-driven correctness effects than their academic ability to produce answers in a 1-minute test. Additionally, the correctness effect itself might not be as relevant a measure as raw P300 amplitude and/or latency, which could be measured independently for each condition. In general, the collectively limited understanding of what ERP measures of correctness reflect, at both cognitive and neurobiological levels, has limited the utility of this technique for math cognition research.

4.6 Conclusions and Future Outlook

Despite some past limitations in this research domain, there is every reason to be optimistic about the future use of ERPs in studies of arithmetic processing. There has been a remarkable consistency in waveform morphology across studies when explicitly looking at number format. The P300 complex response pattern can be seen in every basic plot comparing the response to correct and incorrect answers since the initial studies of arithmetic processing with Arabic numerals were published, although the effects were not always labeled as such (Dickson & Federmeier, 2017; Jost et al., 2004; Niedeggen & Rösler, 1999; Niedeggen et al., 1999; Jaskinski & Coch, 2012; Szűcs & Csépe, 2004; Szűcs & Soltész, 2010). When presenting arithmetic expressions using number words (written or auditory), this response pattern readily transforms to include modulations on the N400.

We find it particularly encouraging that auditory number words elicit a clear correctness effect because this provides experimenters with an online measure of mathematical ability that is independent of the participant's word reading ability. This new approach may be especially useful in the developmental field for researchers who are interested in examining the relationship between acquisition of reading and arithmetic skills.

Acknowledgments

This study was funded by awards R03HD060756 and R21HD079884 from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. This work received computational support from the UTSA Computational System Biology Core, funded by the National Institute on Minority Health and Health Disparities (G12MD007591). This study was conducted as part of the undergraduate honors theses of RB and AR (Experiment 1) and RC (Experiment 2) at the University of Texas at San Antonio. We thank Amanda Martinez-Lincoln and Dr. Ann Eisenberg for their mentorship and contributions to their theses. We also thank Curtiss Chapman and Jacob Momsen for technical assistance with the sound files, and Maria Estefania Cisneros, the voice of the number words.

References

- Avancini C, Soltész F, Szűcs D. Separating stages of arithmetic verification: an ERP study with a novel paradigm. *Neuropsychologia*. 2015; 75:322–329. DOI: 10.1016/j.neuropsychologia.2015.06.016 [PubMed: 26116908]
- Bassok M, Fisher K, Osterhout L. Conceptual integration in arithmetic is the same for digits and words: It's the meaning, stupid!. In: Taatgen NA, Van Rijn H, Schomaker LB, Nerbonne J, editors *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX:

⁵As part of a broader ongoing project, we have also been examining relationships between standardized cognitive measures (i.e., speeded answer production) and ERP effects in arithmetic and have not found consistent/persuasive correlations. In contrast, raw amplitude measures appear more fruitful.

Cognitive Science Society; 2009. 2142–2147. Retrieved from <https://escholarship.org/uc/item/6h32m8mp>

- Bernardo ABI. Language and mathematical problem solving among bilinguals. *The Journal of Psychology*. 2002; 136(3):283–297. DOI: 10.1080/00223980209604156 [PubMed: 12206277]
- Besson M, Macar F. An event-related potential analysis of incongruity in music and other non-linguistic contexts. *Psychophysiology*. 1987; 24(1):14–25. DOI: 10.1111/j.1469-8986.1987.tb01853.x [PubMed: 3575590]
- Brang D, Edwards L, Ramachandran VS, Coulson S. Is the sky 2? Contextual priming in grapheme-color synaesthesia. *Psychological Science*. 2008; 19(5):421–428. DOI: 10.1111/j.1467-9280.2008.02103.x [PubMed: 18466400]
- Campbell JI. Network interference and mental multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1987a; 13(1):109–123. DOI: 10.1037/0278-7393.13.1.109
- Campbell JI. Production, verification, and priming of multiplication facts. *Memory & Cognition*. 1987b; 15(4):349–364. DOI: 10.3758/BF03197037 [PubMed: 3670055]
- Campbell JI, Epp LJ. An encoding-complex approach to numerical cognition in Chinese-English bilinguals. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*. 2004; 58(4):229–244. DOI: 10.1037/h0087447 [PubMed: 15648727]
- Campbell JI, Graham DJ. Mental multiplication skill: Structure, process, and acquisition. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*. 1985; 39(2):338–366. DOI: 10.1037/h0080065
- Campbell JI, Xue Q. Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General*. 2001; 130(2):299–315. DOI: 10.1037/0096-3445.130.2.299 [PubMed: 11409105]
- Connolly JF, Stewart SH, Phillips NA. The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language*. 1990; 39(2):302–381. DOI: 10.1016/0093-934X(90)90016-A [PubMed: 2224497]
- Daltrozzo J, Schön D. Conceptual processing in music as revealed by N400 effects on words and musical targets. *Journal of Cognitive Neuroscience*. 2009; 21(10):1882–1892. DOI: 10.1162/jocn.2009.21113 [PubMed: 18823240]
- Deacon D, Shelley-Tremblay J. How automatically is meaning accessed: a review of the effects of attention on semantic processing. *Frontiers in Bioscience*. 2000; 5(Part E):82–94. DOI: 10.2741/A569
- Dehaene S. Varieties of numerical abilities. *Cognition*. 1992; 44(1):1–42. DOI: 10.1016/0010-0277(92)90049-N [PubMed: 1511583]
- DeLong KA, Urbach TP, Kutas M. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*. 2005; 8(8):1117–1121. DOI: 10.1038/nn1504 [PubMed: 16007080]
- Dickson DS, Federmeier KD. The language of arithmetic across the hemispheres: An event-related potential investigation. *Brain Research*. 2017; 1662:46–56. DOI: 10.1016/j.brainres.2017.02.019 [PubMed: 28237544]
- Domahs F, Domahs U, Schleuisky M, Ratnckx E, Verguts T, Willmes K, Nuerk HC. Neighborhood consistency in mental arithmetic: Behavioral and ERP evidence. *Behavioral and Brain Functions*. 2007; 3(1):66.doi: 10.1186/1744-9081-3-66 [PubMed: 18163911]
- Donchin E, Coles MG. Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*. 1988; 11(3):357–374. DOI: 10.1017/S0140525X00058027
- Duncan-Johnson CC, Donchin E. On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*. 1977; 14(5):456–467. DOI: 10.1111/j.1469-8986.1977.tb01312.x [PubMed: 905483]
- Federmeier KD, Kutas M. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*. 1999a; 41(4):469–495. DOI: 10.1006/jmla.1999.2660
- Federmeier KD, Kutas M. Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research*. 1999b; 8(3):373–392. DOI: 10.1016/S0926-6410(99)00036-1 [PubMed: 10556614]

- Federmeier KD, Kutas M. Meaning and modality: Influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2001; 27(1):202–224. DOI: 10.1037/0278-7393.27.1.202
- Federmeier KD, Kutas M, Dickson DS. A common neural progression to meaning in about a third of a second. In: Hickok GS, Small SL, editors *Neurobiology of Language*. Holland: Elsevier; 2015. 557–567.
- Federmeier KD, McLennan DB, De Ochoa E, Kutas M. The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*. 2002; 39(2):133–146. DOI: 10.1111/1469-8986.3920133 [PubMed: 12212662]
- Fisher K, Bassok M, Osterhout L. When two plus two does not equal four: Event-related potential responses to semantically incongruous arithmetic word problems. In: Ohlsson S, Catrambone R, editors *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society; 2010. 1571–1576. Retrieved from <https://escholarship.org/uc/item/6h32m8mp>
- French-Mestre C, Vaid J. Activation of number facts in bilinguals. *Memory & Cognition*. 1993; 21(6): 809–818. DOI: 10.3758/BF03202748 [PubMed: 8289658]
- Galfano G, Mazza V, Angrilli A, Umiltà C. Electrophysiological correlates of stimulus-driven multiplication facts retrieval. *Neuropsychologia*. 2004; 42(10):1370–1382. DOI: 10.1016/j.neuropsychologia.2004.02.010 [PubMed: 15193945]
- Galfano G, Penolazzi B, Vervaeck I, Angrilli A, Umiltà C. Event-related brain potentials uncover activation dynamics in the lexicon of multiplication facts. *Cortex*. 2009; 45(10):1167–1177. DOI: 10.1016/j.cortex.2008.09.003 [PubMed: 18976746]
- Geyer A, Holcomb P, Kuperberg G, Perlmutter N. Plausibility and sentence comprehension. An ERP study. *Cogn Neurosci Suppl*. 2006 Abstract.
- Guthormsen AM, Fisher KJ, Bassok M, Osterhout L, DeWolf M, Holyoak KJ. Conceptual integration of arithmetic operations with real world-knowledge: Evidence from event-related potentials. *Cognitive Science*. 2016; 40(3):723–757. DOI: 10.1111/cogs.12238 [PubMed: 25864403]
- Holcomb PJ, Neville HJ. Natural speech processing: An analysis using event-related brain potentials. *Psychobiology*. 1991; 19(4):286–300. DOI: 10.3758/BF03332082
- Jasinski EC, Coch D. ERPs across arithmetic operations in a delayed answer verification task. *Psychophysiology*. 2012; 49(7):943–958. DOI: 10.1111/j.1469-8986.2012.01378.x [PubMed: 22563982]
- Johnson R, Donchin E. P300 and stimulus categorization: Two plus one is not so different from one plus one. *Psychophysiology*. 1980; 17(2):167–178. DOI: 10.1111/j.1469-8986.1980.tb00131.x [PubMed: 7375618]
- Jost K, Hennighausen E, Rösler F. Comparing arithmetic and semantic fact retrieval: Effects of problem size and sentence constraint on event-related brain potentials. *Psychophysiology*. 2004; 41(1):46–59. DOI: 10.1111/1469-8986.00119 [PubMed: 14693000]
- Kluender R, Kutas M. Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*. 1993; 5(2):196–214. DOI: 10.1162/jocn.1993.5.2.196 [PubMed: 23972154]
- Koelsch S, Kasper E, Sammler D, Schulze K, Gunter T, Friederici AD. Music, language and meaning: signatures of semantic processing. *Nature Neuroscience*. 2004; 3(7):302–307. DOI: 10.1038/nrn1197
- Kolk HH, Chwilla DJ, Van Herten M, Oor PJ. Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*. 2003; 85(1):1–36. DOI: 10.1016/S0093-934X(02)00548-5 [PubMed: 12681346]
- Kutas M. Event-related brain potentials (ERPs) elicited during rapid serial visual presentation of congruous and incongruous sentences. *Electroencephalography and Clinical Neurophysiology*. 1987; 40(Supplement):406–411.

- Kutas M, Federmeier KD. Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*. 2000; 4(12):463–470. DOI: 10.1038/307161a0 [PubMed: 11115760]
- Kutas M, Federmeier KD. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*. 2011; 62:621–647. DOI: 10.1146/annurev.psych.093008.131123
- Kutas M, Hillyard SA. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*. 1980; 207(4427):203–205. DOI: 10.1126/science.7350657 [PubMed: 7350657]
- Kutas M, Hillyard SA. Brain potentials during reading reflect word expectancy and semantic association. *Nature*. 1984; 307(5947):161–163. [PubMed: 6690995]
- Lawrence MA. ez: Easy analysis and visualization of factorial experiments. 2013. R package version 4.4-0[Software]. Available from <https://CRAN.R-project.org/package=ez>
- Lee CL, Federmeier KD. Wave-ering: An ERP study of syntactic and semantic context effects on ambiguity resolution for noun/verb homographs. *Journal of Memory and Language*. 2009; 61(4): 538–555. DOI: 10.1016/j.jml.2009.08.003 [PubMed: 20161361]
- Lopez-Calderon J, Luck SJ. ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*. 2014; 8:213.doi: 10.3389/fnhum.2014.00213 [PubMed: 24782741]
- Marian V, Blumenfeld HK, Kaushanskaya M. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*. 2007; 4(50):940–967. DOI: 10.1044/1092-4388(2007/067)
- Martinez-Lincoln A, Cortinas C, Wicha NY. Arithmetic memory networks established in childhood are changed by experience in adulthood. *Neuroscience Letters*. 2015; 584:325–330. DOI: 10.1016/j.neulet.2014.11.010 [PubMed: 25445361]
- McLaughlin J, Tanner D, Pitkänen I, Frenck-Mestre C, Inoue K, Valentine G, Osterhout L. Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*. 2010; 60(s2): 123–150. DOI: 10.1111/j.1467-9922.2010.00604.x
- Moreno EM, Kutas M. Processing semantic anomalies in two languages: An electrophysiological exploration in both languages of Spanish–English bilinguals. *Cognitive Brain Research*. 2005; 22(2):205–220. DOI: 10.1016/j.cogbrainres.2004.08.010 [PubMed: 15653294]
- Niedeggen M, Rösler F. N400 effects reflect activation spread during retrieval of arithmetic facts. *Psychological Science*. 1999; 10(3):271–276. DOI: 10.1111/1467-9280.00149
- Niedeggen M, Rösler F, Jost K. Processing of incongruous mental calculation problems: Evidence for an arithmetic N400 effect. *Psychophysiology*. 1999; 36(3):307–324. DOI: 10.1017/S0048577299980149 [PubMed: 10352554]
- Nieuwenhuis S, Aston-Jones G, Cohen JD. Decision making, the P3, and the locus coeruleus—norepinephrine system. *Psychological Bulletin*. 2005; 131(4):510–532. DOI: 10.1037/0033-2909.131.4.510 [PubMed: 16060800]
- Núñez-Peña MI, Escera C. An event-related brain potential study of the arithmetic split effect. *International Journal of Psychophysiology*. 2007; 64(2):165–173. DOI: 10.1016/j.ijpsycho.2007.01.007 [PubMed: 17360062]
- Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*. 1971; 9(1):97–113. DOI: 10.1016/0028-3932(71)90067-4 [PubMed: 5146491]
- Osterhout L, McLaughlin J, Pitkänen I, Frenck-Mestre C, Molinaro N. Novice learners, longitudinal designs, and event-related potentials: A means for exploring the neurocognition of second language processing. *Language Learning*. 2006; 56(s1):199–230. DOI: 10.1111/j.1467-9922.2006.00361.x
- Paller KA, McCarthy G, Wood CC. Event related-potentials elicited by deviant endings to melodies. *Psychophysiology*. 1992; 29(2):202–206. <https://doi.org/10.1111/j.1469-8986.1992.tb01686.x>. [PubMed: 1635962]
- Polich J. Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology*. 2007; 118(10):2128–2148. DOI: 10.1016/j.clinph.2007.04.019 [PubMed: 17573239]

- Polich J. Neuropsychology of P300. In: Luck SJ, Kappenman ES, editors Oxford Handbook of Event-Related Potential Components. Oxford, U.K.: Oxford Univ. Press; 2012. 159–188.
- Polich J, Criado JR. Neuropsychology and neuropharmacology of P3a and P3b. *International Journal of Psychophysiology*. 2006; 60(2):172–185. DOI: 10.1016/j.ijpsycho.2005.12.012 [PubMed: 16510201]
- R Core Team. R Foundation for Statistical Computing (Version 3.3R: A language and environment for statistical computing2) [Software]. Vienna: 2015. R: A language and environment for statistical computing. Available from <https://www.R-project.org>
- Ruchkin DS, Johnson R, Grafman J, Canoune H, Ritter W. Distinctions and similarities among working memory processes: An event-related potential study. *Cognitive Brain Research*. 1992; 1(1):53–66. DOI: 10.1016/0926-6410(92)90005-C [PubMed: 15497435]
- Salillas E, Wicha NY. Early learning shapes the memory networks for arithmetic: evidence from brain potentials in bilinguals. *Psychological Science*. 2012; 23(7):745–755. DOI: 10.1177/0956797612446347 [PubMed: 22707225]
- Scudder MR, Federmeier KD, Raine LB, Direito A, Boyd JK, Hillman CH. The association between aerobic fitness and language processing in children: Implications for academic achievement. *Brain and Cognition*. 2014; 87:140–152. DOI: 10.1016/j.bandc.2014.03.016 [PubMed: 24747513]
- Szcs D, Csépe V. Access to numerical information is dependent on the modality of stimulus presentation in mental addition: A combined ERP and behavioral study. *Cognitive Brain Research*. 2004; 19(1):10–27. DOI: 10.1016/j.cogbrainres.2003.11.002 [PubMed: 14972354]
- Szcs D, Soltész F. Event-related brain potentials to violations of arithmetic syntax represented by place value structure. *Biological Psychology*. 2010; 84(2):354–367. DOI: 10.1016/j.biopsycho.2010.04.002 [PubMed: 20394796]
- Tanner D, McLaughlin J, Herschensohn J, Osterhout L. Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition*. 2013; 16(2):367–382. DOI: 10.1017/S1366728912000302
- Tanner D, Morgan-Short K, Luck SJ. How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*. 2015; 52(8):997–1009. DOI: 10.1111/psyp.12437 [PubMed: 25903295]
- Van Berkum JJ, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2005; 31(3):443–467. DOI: 10.1037/0278-7393.31.3.443
- Van Petten C, Luka BJ. Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*. 2012; 83(2):176–190. DOI: 10.1016/j.ijpsycho.2011.09.015 [PubMed: 22019481]
- Voss JL, Schendan HE, Paller KA. Finding meaning in novel geometric shapes influences electrophysiological correlates of repetition and dissociates perceptual and conceptual priming. *NeuroImage*. 2010; 49(3):2879–2889. DOI: 10.1016/j.neuroimage.2009.09.012 [PubMed: 19770053]
- Wagner RK, Torgesen JK, Rashotte CA. Comprehensive test of phonological processing (CTOPP). Austin, TX: Pro-Ed; 1999.
- Wechsler D. Wechsler Individual Achievement Test. San Antonio, TX: Psychological Corporation; 1992.
- Wicha NY, Bates EA, Moreno EM, Kutas M. Potato not Pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*. 2003; 346(3):165–168. DOI: 10.1016/S0304-3940(03)00599-8 [PubMed: 12853110]
- Wicha NY, Moreno EM, Kutas M. Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex*. 2003; 39(3):483–508. DOI: 10.1016/S0010-9452(08)70260-0 [PubMed: 12870823]
- Wicha NYY, Moreno EM, Kutas M. Anticipating words and their gender: An event-related potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*. 2004; 16(7):1272–1288. DOI: 10.1162/0898929041920487 [PubMed: 15453979]

- Wlotko EW, Federmeier KD. So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*. 2012; 62(1): 356–366. DOI: 10.1016/j.neuroimage.2012.04.054 [PubMed: 22565202]
- Wlotko EW, Federmeier KD. Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*. 2015; 68:20–32. DOI: 10.1016/j.cortex.2015.03.014 [PubMed: 25987437]
- Wlotko EW, Lee CL, Federmeier KD. Language of the aging brain: Event-related potential studies of comprehension in older adults. *Language and Linguistics Compass*. 2010; 4(8):623–638. DOI: 10.1111/j.1749-818X.2010.00224.x [PubMed: 20823949]
- Woodcock RW, McGrew KS, Mather N. Woodcock-Johnson III Tests of Achievement. Itasca, IL: Riverside Publishing; 2001a.
- Woodcock RW, McGrew KS, Mather N. Woodcock-Johnson III Tests of Cognitive Abilities. Itasca, IL: Riverside Publishing; 2001b. 371–401.
- Woodcock RW, McGrew KS, Mather N, Schrank FA. Woodcock-Johnson III Diagnostic Supplement to the Tests of Cognitive Abilities. Itasca, IL: Riverside Publishing; 2007.

Digit Task

Auditory Task

Figure 1.

Trial structure in Experiment 1 for the task with digit operands (top) and for the task with spoken number word operands (bottom), from left to right across time.

Figure 2.

Layout of electrode channels used in all recordings, oriented with the front of the head at the top of the figure. Darkened circles indicate electrodes utilized in distributional analyses.

Figure 3.

Grand average ERPs measured to the onset of correct (black) and incorrect (red) solutions with a 100ms prestimulus baseline, at representative channels indicated on the scalp map with an X. Negative voltage is plotted up. Solutions preceded by Arabic digit operands are on the left and solutions preceded by spoken number words are on the right. A robust correctness effect is visible, featuring an N400 in the auditory context and a P300 in the digit context. At top, topographic maps of the correctness effect are plotted for digits and auditory formats using the analysis time windows for the P300 or N400, respectively. Note that for

illustration purposes and consistency with our discussion we show the digit format effect as a positivity (correct minus incorrect solutions) and the auditory format effect as a negativity (incorrect minus correct solutions).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Difference ERPs (Correct - Incorrect)

Figure 4.

Difference ERPs (correct minus incorrect) are plotted for each operand context (auditory in black; digit in red). The time windows used for analyses of the N400/P300 and subsequent effects are indicated by their location above the plot and are similarly color-coded for each trial type. The late effect for auditory trials did not reach significance.

Figure 5.

Trial structure in Experiment 2 for the task with auditory operands, from left to right across time. The critical manipulation is the length of delay after the second operand.

Figure 6.

Grand average ERPs measured from the onset of correct (black) and incorrect (red) solutions with a 100ms prestimulus baseline, from representative channels indicated on the scalp map with an X. At left are the responses to solutions when there was a short delay (150ms) following the second operand. Corresponding plots for the response to solutions when there was a longer delay (1000ms) are plotted on the right. A robust N400 correctness effect is visible with both delays. At top, topographic maps of the correctness effect (incorrect minus correct solutions) are plotted using the analysis time window for the N400.

Figure 7.

Grand average ERPs at a representative middle central electrode (MiCe) contrast the waveform componentry and timing across all experiments and critical conditions.

Table 1
Participants' offline performance on cognitive measures

	<u>Mean</u>	<u>SE</u>
<u>Experiment 1</u>		
Math Fluency ^a	98.35	2.48
Working Memory ^b	98.31	2.74
Phonological Awareness ^c	102.77	2.09
<u>Experiment 2</u>		
Math Fluency ^d	99.50	2.48
Working Memory ^b	100.35	3.07
Phonological Awareness ^c	100.42	1.97

Note. $N = 26$ in each experiment. All mean scores are based on standardized measures where 100 is the age-based norm and 15 points reflects 1 SD outside the norm.

^a Math fluency was measured by the WJ-III, which is a combination of multiplication, addition, and subtraction problems.

^b Working memory was measured by the Numbers Reversed task on the WJ-III.

^c Phonological awareness was measured by the composite score of the Elision and Blending Words subtests of the CTOPP.

^d Math fluency is reported from performance on the multiplication subtest of the WIAT.