COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 40 (2016) 723–757

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12238

# Conceptual Integration of Arithmetic Operations With Real-World Knowledge: Evidence From Event-Related Potentials

Amy M. Guthormsen,[a] Kristie J. Fisher,[b] Miriam Bassok,[c] Lee Osterhout,[c] Melissa DeWolf,[d] Keith J. Holyoak[d]

[a]*Los Alamos National Laboratory*
[b]*Microsoft Studios*
[c]*University of Washington*
[d]*University of California, Los Angeles*

Received 23 May 2014; received in revised form 5 January 2015; accepted 23 January 2015

**Abstract**

Research on language processing has shown that the disruption of conceptual integration gives rise to specific patterns of event-related brain potentials (ERPs)—N400 and P600 effects. Here, we report similar ERP effects when adults performed cross-domain conceptual integration of analogous semantic and mathematical relations. In a problem-solving task, when participants generated labeled answers to semantically aligned and misaligned arithmetic problems (e.g., *6 roses* + *2 tulips* = *?* vs. *6 roses* + *2 vases* = *?*), the second object label in misaligned problems yielded an N400 effect for addition (but not division) problems. In a verification task, when participants judged arithmetically correct but semantically misaligned problem sentences to be "unacceptable," the second object label in misaligned sentences elicited a P600 effect. Thus, depending on task constraints, misaligned problems can show either of two ERP signatures of conceptual disruption. These results show that well-educated adults can integrate mathematical and semantic relations on the rapid timescale of within-domain ERP effects by a process akin to analogical mapping.

*Keywords:* Analogical mapping; Mathematical reasoning; Semantic alignment; ERP; N400 effect; P600 effect

## 1. Introduction

Mathematics is a domain that requires manipulation of abstract symbols in accordance with formal rules. A common view equates mathematical thinking with the formal

Correspondence should be sent to Miriam Bassok, Department of Psychology, University of Washington, Box 35125, Seattle, WA 98195. E-mail: mbassok@u.washington.edu

properties of mathematics, leading to the assumption that people's knowledge of numbers and calculations is separated from the rest of their conceptual knowledge. An impressive body of research on mathematical cognition can be interpreted as providing support for the view that mathematical knowledge is to some extent isolable. For example, researchers have identified specific brain areas that are apparently devoted to numerical representations and calculation (for a review see Dehaene, Molko, Cohen, & Wilson, 2004). Evidence from patients with neurological deficits indicates that calculation abilities can be preserved despite severe impairments in language capabilities (Pesenti, Thioux, Seron, & De Volder, 2000; Varley, Klessinger, Romanowski, & Siegal, 2005; Zago et al., 2001). Researchers who examine how people retrieve arithmetic facts (e.g., $3 + 4 \underline{=} 7$) have posited the existence of specialized number networks that are analogous to, but presumably separate from, networks of semantic knowledge (for reviews see Ashcraft, 1992; Campbell, 1995; for an alternative perspective see Campbell & Metcalfe, 2009; Campbell & Sacher, 2012). Furthermore, it appears that mathematical cognition develops from an innate "number sense" that is present at birth (Barth, La Mont, Lipton, & Spelke, 2005).

### *1.1. Semantic alignment in mathematical modeling*

The view that mathematics is an isolable knowledge domain is supported mostly by studies that employed tasks involving the abstract and formal properties of mathematics. However, the adaptive value of numbers and calculations lies in their role as tools for solving real-world problems. The usefulness of these tools depends crucially on their fit to the situations in which they are applied. For example, most people know that $1 + 1 \underline{=} 2$. If John has one son and his wife delivers another son, we readily compute that he now has two sons. However, if John had one wife and now has another wife, most likely he now has only one wife. In the latter case, conceptual knowledge leads us to refrain from applying addition. More generally, because people have to decide whether and when to apply their mathematical knowledge (here, whether or not to use addition), they must coordinate it with their conceptual knowledge. Thus, knowledge about when and how to apply mathematical procedures is guided by conceptual understanding.

The process by which people reason with mathematical representations of real-world situations is referred to as *mathematical modeling*.[1] Mathematical modeling is a complex cognitive process that requires the problem solver to select the appropriate mathematical operations to perform based on a description of a real-world situation (Kintsch & Greeno, 1985). Because this process may be quite effortful, people often circumvent it by relying on various shortcut heuristics (e.g., Clement, Lochhead, & Monk, 1981; Fisher, Borchert, & Bassok, 2011; Martin & Bassok, 2005). People's tendency to engage in mathematical modeling can be influenced by a variety of factors, such as their mathematical knowledge (Hinsley, Hayes, & Simon, 1977), the problem's mathematical format (Fisher et al., 2011; Novick, 1990), or the causal relations in the real-world situation being described (Mochon & Sloman, 2004).

When people do engage in mathematical modeling, they are sensitive to *semantic alignment*—a preference for sensible analogical mappings between mathematical and semantic relations (for a review see Bassok, 2001). Bassok, Chase, and Martin (1998) have shown such semantic alignments for the arithmetic operations of addition and division (see Fig. 1). They found that people align categorically related objects (co-hyponyms) with the symmetric roles of addends (e.g., *roses* + *tulips*; *priests* + *ministers*), and align functionally related objects with the asymmetric roles of dividend and divisor (e.g., *tulips* ÷ *vases*; *priests* ÷ *parishioners*). Such semantic alignments result in pragmatically sensible mathematical models, whereas reversing the correspondences (i.e., addition of functionally related entities or division of categorically related entities) creates non-sensible models. To illustrate, whereas (contrary to an old adage) it makes sense to add apples and oranges, in most situations it does not make sense to add apples and baskets. Similarly, while it makes sense to divide apples among baskets, the meaning of dividing apples among oranges is much less obvious.

Evidence also indicates that, at least for addition facts, semantic alignments are highly automatic (Bassok, Pedigo, & Oskarsson, 2008). In a digit-verification task (e.g., *Did you see 3? YES/NO*), addition facts (e.g., *3 + 5*) were primed with pairs of object sets having a categorical semantic relation (aligned with addition; e.g., *tulips, daisies*), a functional relation (misaligned with addition; e.g., *tulips, vases*), or no relation (e.g., *clocks, chickens*). Participants in this type of task typically exhibit a "sum effect" (Lefevre, Bisanz, & Mrkonjic, 1988), taking longer to reject the sum as having *not* been present in the initial digit pair (e.g., *Did you see 8? Answer: NO*) relative to a foil (e.g., *Did you see 9?*

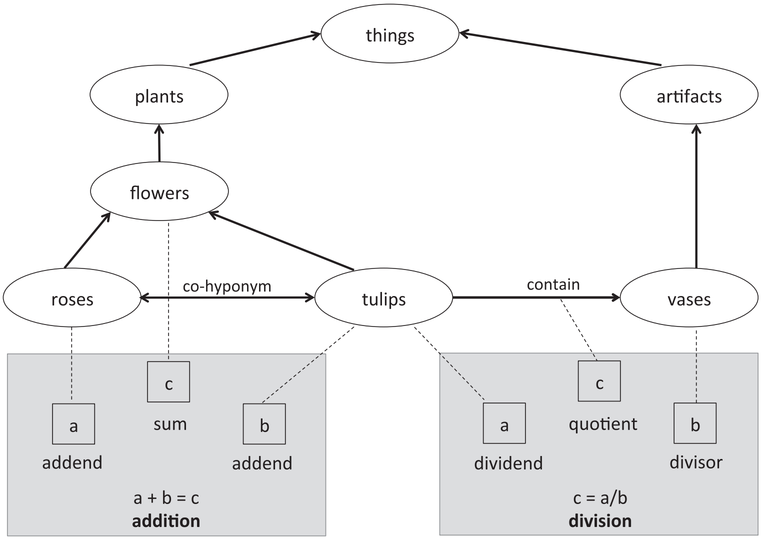


Fig. 1. Semantic alignments for addition and division operators. Conceptually meaningful sets are bound into mathematical roles (dashed lines) that serve as inputs to addition (addends) and division operators (dividend, divisor). The symmetric roles for addition map naturally onto the symmetric conceptual relation of *co-hyponym* (e.g., *roses* and *tulips*); the asymmetric roles for division map onto asymmetric conceptual relations such as *contains* (e.g., *tulips* and *vases*). Unlabeled vertical arrows represent the relation of category membership ("is a").

*Answer: NO*). Addition facts primed with categorical semantic relations, which are aligned with addition, showed the sum effect. However, the sum effect was *not* obtained when misaligned or unrelated object pairs were used as primes.

The finding of Bassok et al. (2008) suggests that the alignment of arithmetic and semantic relations could be a special case of relational priming (Spellman, Holyoak, & Morrison, 2001). Though relational priming has generally been found only when participants were instructed to attend to relations, it appears that the coordination of the categorical relation with the addition operator is overlearned due to implicit and explicit instruction. In less familiar tasks (e.g., algebraic modeling), in which alignments between semantic and arithmetic relations are less well established, only those people who engage in a strategy of mathematical modeling show sensitivity to semantic alignment (Fisher & Bassok, 2009).

### *1.2. Using ERP methods to assess conceptual integration*

The hypothesis that mathematical modeling is guided by semantic alignment implies that linguistic and mathematical knowledge need to be integrated to create a meaningful whole, while maintaining consistency with relevant contextual factors. Such *conceptual integration* is intrinsic to the process of language comprehension, which requires the integration of consecutive words into meaningful sentences. Similarly, the comprehension of arithmetic problems requires the integration of digits and operator symbols into a coherent arithmetic equation.

Conceptual integration in both the language and mathematical domains has been studied using event-related brain potentials (ERPs). ERPs reflect the summed, simultaneously occurring electrical activity in the brain that occurs following some specific eliciting event. ERPs are derived from the ongoing EEG by time-locking to the onset of each critical word in a sentence (for example, a semantically anomalous word) and extracting a second or two of EEG activity. These segments of EEG are then averaged together to extract the ERP. ERPs are typically described in terms of latency (in milliseconds) and polarity (positive- or negative-going). ERP methods are well suited to study conceptual integration because they allow for the comparison of the brain's electrical responses to individual, sequentially presented items as a person attempts to integrate them into a meaningful whole. Extensive work on language processing has established that conceptual integration is disrupted when violations of meaning or structure are present within the items to be integrated. Moreover, different types of violations elicit distinctive ERP responses. Anomalies involving semantic meaning (e.g., the word BAKE in *The cat will BAKE the food I leave on the porch*) elicit a larger-amplitude N400 component, compared to a sentence in which the word is semantically appropriate. N400 is a negative-going wave that peaks at about 400 ms after presentation of the anomalous word (e.g., Kutas & Federmeier, 2011; Kutas & Hillyard, 1980, 1984; Osterhout & Nicol, 1999). By contrast, a variety of syntactic anomalies (including anomalies of phrase structure and morphosyntax; e.g., *The cat will EATING the food I leave on the porch*) elicit a large centro-parietal positive wave that starts at about 500 ms and persists for at least half a second (the *P600 effect*; Osterhout & Holcomb, 1992, 1995).

The amplitude of both of these ERP components is modulated by the degree to which a violation disrupts conceptual or grammatical integration (Kutas & Federmeier, 2000; Osterhout, Holcomb, & Swinney, 1994). For example, the N400 effect elicited by *He takes sugar and cream in his JUICE*, relative to *COFFEE*, is likely to be smaller in amplitude than that elicited by *He takes sugar and cream in his SOCKS* (Kutas & Hillyard, 1984). Federmeier and Kutas (1999, 2001) found that the strength of the N400 effect varies with the degree of unexpectedness of the violated word. Within-category violations generate a stronger N400 effect than expected exemplars, and between-category violations (or unexpected items from a different semantic category) generate the strongest N400 response. Furthermore, Osterhout and Mobley (1995) found that when there is *any* kind of violation within a sentence (either syntactic or semantic), an N400-like effect is also elicited by the final word in the sentence, even when that word is both semantically and syntactically appropriate (see Kutas & Federmeier, 2011, for a review). This "last-item" N400 effect is likely related to the cognitive requirements of the *delayed verification* experimental paradigm typically used in language research. Participants are asked to make a binary judgment about the "acceptability" of the sentence they just saw (usually they are not instructed to look for any particular type of error). Thus, when participants reach the end of a sentence that contained a violation, the entire sentence must now be categorized as "unacceptable." The N400 effect to the final word in the sentence may be a result of this judgment processing.

The dichotomy between semantic processing (N400) versus syntactic processing (P600) generalizes well across languages and stimuli (e.g., a P600 effect has been observed for syntactic anomalies involving phrase structure, verb subcategorization, verb tense, subject-verb number agreement, number and gender pronoun-antecedent agreement, case, and constituent movement; see Osterhout & Nicol, 1999). The picture is more complex for anomalies based on semantic verb-argument violations, which seem to have both grammatical and semantic qualities (e.g., Kim & Osterhout, 2005; for a review see Kuperberg, 2007). Incongruent metaphors have been reported to elicit either N400 or P600 effects (see Yang, Bradley, Huq, Wu, & Krawczyk, 2013).

Similar ERP effects have also been observed in experiments involving non-linguistic tasks. For example, semantically implausible sequences of events depicted in videos (Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008; Sitnikova, Kuperberg, & Holcomb, 2003) or pictures (West & Holcomb, 2002) elicit N400 effects. Võ and Wolfe (2013) found that for visual scenes, semantically inconsistent objects elicited an N400 effect, whereas mild (but not extreme) violations of expected object location elicited a P600 effect. P600-like effects are also elicited by violations of harmonic scale progression (e.g., Patel & Daniele, 2002; Patel, Gibson, Ratner, Besson, & Holcomb, 1998).

Most relevant to the current study are reports of similar ERP effects in the context of arithmetic problem solving. Mathematically incorrect answers to addition (Szucs & Csépe, 2004, 2005), multiplication (Jost, Henninghausen, & Rosler, 2004; Niedeggen & Rösler, 1999; Niedeggen, Rosler, & Jost, 1999), and subtraction and division problems (Wang, Kong, Tang, Zhuang, & Li, 2000) also elicit an N400-like effects. This effect has sometimes been shown to occur earlier than the N400 effect observed in sentence processing, with a peak occurring as early as 270 ms after stimulus presentation (as opposed to 400 ms, which is typically observed for linguistic stimuli; Wang et al., 2000). Similarly, rule violations in arithmetic problems elicit a P600-like positivity. For instance, violations of the syntax of arithmetic operators (e.g., *10 − * = 5*) elicit a P600-like effect (Martín-Loeches, Casado, Ganzalo, De Heras, & Fernández-Frías, 2006). Also, an arithmetic series in which the final term is inconsistent with the relational pattern established by the earlier terms elicits a P600 effect (e.g., 7, 10, 13, 16, 19, 22, 50), with the size of the effect varying increasing with the magnitude of the violation (Núñez-Peña & Honrubia-Serrano, 2004).

### 1.3. Overview of the present experiments

Though the process of conceptual integration in arithmetic appears to be similar in nature to that operating in language processing, it is unknown how conceptual integration proceeds when information must be integrated *across* the domains of semantic and arithmetic knowledge, as is the case when people perform mathematical modeling. As we mentioned earlier, people have highly systematic expectations for what object relations should be included in different arithmetic problems (Bassok et al., 1998; Martin & Bassok, 2005). Moreover, for addition facts, the coordination of semantic and arithmetic relations can occur as fluently as arithmetic fact retrieval (Bassok et al., 2008). The strength and fluency of expectations for semantic alignment suggest that ERP methodology could capture violations of semantic alignment in arithmetic word problems.

Here, we report experiments in which ERPs were recorded while participants processed mathematical problem statements in which the numerical operands bore object labels, a task likely to evoke integration across arithmetic and semantic domains for participants who engage in mathematical modeling. ERPs were extracted from the ongoing EEG by time-locking to a critical event in each trial (e.g., the sum in an equation) and extracting a second or two of contiguous EEG signal. The time-locked segments of EEG were then averaged over all trials of that type. As discussed above, Bassok et al. (1998) have shown that specific alignments arise when addition is paired with co-hyponyms (categorical relations) and when division is paired with words instantiating the "contains" relation (functional relations). We selected these two types of arithmetic problems in order to evaluate integration of semantic and arithmetic relations. Our basic aim was to apply ERP methods to assess semantic and structural influences of conflict created by misaligning object relations and arguments of arithmetic operations. To achieve this aim, we used two distinct paradigms, one requiring generation of a semantically meaningful answer to aligned and misaligned problems (Experiment 1), and one requiring an acceptability judgment for aligned and misaligned problem sentences (Experiment 2). As we elaborate in the introductions to the two experiments, we hypothesized that the generation task would focus participants' attention on semantic violations in misaligned problems (yielding an N400 effect), whereas the acceptability judgment task would focus participants' attention on structural violations (yielding a P600 effect).

## 2. Experiment 1

Experiment 1 was designed to introduce potential conflict in coordinating the conceptual domains of arithmetic and semantic knowledge. College students generated solutions to arithmetic problems in which the numerical operands bore object labels, a task that might be expected to encourage integration across arithmetic and semantic relations (Bassok et al., 1998). In Experiment 1A adults had to generate labeled answers to simple addition problems that were either semantically aligned (e.g., *3 tulips* + *5* <u>*roses*</u> = ***8 flowers***) or misaligned (e.g., *3 tulips* + *5* <u>*vases*</u> = ***8 things in a flower shop***). Because the problem-solving task required generating semantically appropriate labels for the numerical answers, we hypothesized that the second object label in misaligned problems would be interpreted as a semantic violation, yielding an N400 effect. Experiment 1B extended this paradigm to division problems (e.g., *12 tulips/3* <u>*vases*</u> = ***4 tulips per vase***, or *12 tulips/3* <u>*roses*</u> = ***4 tulips per rose***).

## 3. Experiment 1A

### 3.1. Method

#### 3.1.1. Participants

Ten women from the University of Washington aged 20–33 years (mean = 26 years) participated as volunteers. All were native English speakers with normal or corrected-to-normal vision.

#### 3.1.2. Stimuli and design

Each participant was asked to generate an answer to 141 arithmetic problems. These consisted of 90 addition problems, half with labels denoting categorically related and half with labels denoting functionally related object sets. The remaining 51 were filler problems involving subtraction. The object-set labels for the target trials were chosen by first generating 45 word triplets to serve as the object-set labels in the target problems. Each word triplet consisted of a common base set (e.g., *tulips*), a categorically related set that formed a co-hyponym (e.g., *roses*), and a functionally related set (e.g., *vases*). Each word triplet served to create one problem in each condition. The categorically related and functionally related words were closely matched in word length (means of 6.3 and 6.4 letters, respectively) and frequency (means of 4,374 and 4,275, respectively, from the CORPORA database; http://www.corpus.byu.edu). To validate our selection of word stimuli, we asked a separate group of 77 participants to rate, on a scale of 1–7, the degree to which the word pairs were categorically related ($n$ = 36) or functionally related ($n$ = 41). Each participant rated half of the categorically related and half of the functionally related word pairs used in the experiment. Their ratings were consistent with our classification, and there was no overlap in the rating distributions for the categorical and the functional word pairs. Specifically, the categorical ratings of the categorically related word pairs ($M$ = 5.83, $SD$ = 0.89) were significantly higher than those of the functionally related word pairs ($M$ = 3.14, $SD$ = 1.28), $t(35) = 9.00$, $p < .001$. Similarly, the functional ratings of the functionally related word pairs ($M$ = 5.57, $SD$ = 1.10) were significantly higher than those of the categorically related word pairs ($M$ = 2.31, $SD$ = 1.26), $t(40) = 9.41$, $p < .001$.

The pairs of numbers that served as addends were selected from the full set of single digits, excluding 0. A representative set of aligned and misaligned addition problems appears in the left column of Table 1.

The subtraction filler problems also had two labeled operands. Although the operand labels were not designed to vary systematically in terms of alignment, we did attempt to generate some variability in the extent to which the operand labels matched the operation. Seventeen subtraction problems had unrelated sets (e.g., *8 filters − 3 stripes*), 17 had set/subset sets (e.g., *9 politicians − 3 senators*), and 17 had one-to-one sets (e.g., *9 typewriters − 2 secretaries*).

The problems were placed in a pseudo-randomized order, so that problems from the same triplet (e.g., *2 tulips + 3 roses* and *5 tulips + 4 vases*) did not appear in the same half of the list, and so that no more than three trials of any type occurred in succession. We utilized this sequence and its inverse, randomly assigning participants to one or the other order of trials.

### 3.1.3. *Procedure*

Participants were tested in one session that lasted between 1.5 and 2.5 h. They were seated in a comfortable chair situated in a sound-attenuating room and were told that their task was to solve simple addition and subtraction problems. Participants were also told that the numbers in each problem would have object labels, and that their task was to generate an answer that included both a numerical component and an object label for it.

The events on a single trial are schematized in Fig. 2. Each trial began with a fixation point. To minimize eye movements, each problem was presented sequentially so that only one word or number appeared centered on the screen at a given time, with a stimulus presentation time of 650 ms for each word or number (and 0 ms delay between successive words). This timing was chosen to correspond to that used in standard ERP sentence-processing experiments (e.g., Osterhout & Holcomb, 1992). At the end of the problem, the

Table 1
Examples of stimuli by condition for Experiments 1A and 1B

| | Operation | |
|---|---|---|
| Alignment | Addition | Division |
| Aligned | 3 chemists + 2 physicists | 12 cars/6 mechanics |
| | 6 cookies + 3 brownies | 21 rubies/3 necklaces |
| Misaligned | 4 cars + 2 mechanics | 12 chemists/3 physicists |
| | 2 rubies + 6 necklaces | 20 cookies/5 brownies |

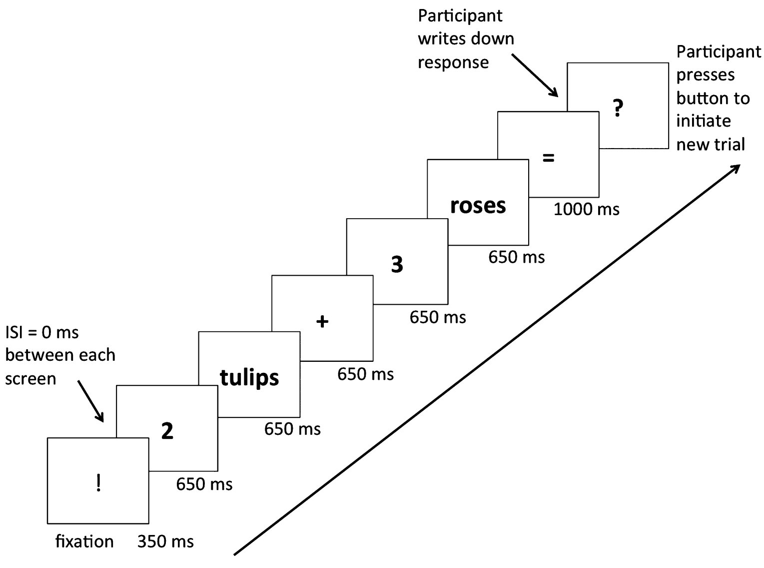


Fig. 2. The sequence of stimuli presented during trials for Experiment 1A. The same sequence was used for Experiment 1B with division instead of addition problems.

participant was free to write down an answer on an answer sheet. When ready, the participant began a new trial by pressing a button on a serial response box.

#### 3.1.4. EEG recordings

Continuous EEG was recorded using tin electrodes attached to an elastic cap (Electro-cap International) in accordance with the extended 10–20 system (Nuwer et al., 1998). Recordings were obtained from homologous positions of the left and right prefrontal (Fp1, Fp2), frontal (F3, F4), inferior frontal (F7, F8), temporal (T7, T8), central (C3, C4), parietal (P3, P4), posterior parietal (P7, P8), and occipital (O1, O2) sites, as well as from three midline locations (Fz, Cz, Pz). Continuous analog-to-digital conversion of the EEG and stimulus trigger codes was performed at a sampling frequency of 200 Hz.

Vertical eye movements and blinks were monitored by means of two electrodes, one placed beneath the left eye and one placed to the right of the right eye. The above 19 channels were referenced to an electrode placed over the left mastoid bone and were amplified with a bandpass of 0.01–100 Hz (3 dB cutoff) by an SAI bioamplifier system. Activity over the right mastoid was recorded on the twentieth channel to determine if there were any effects of the experimental variables on the mastoid recordings. No such effects were observed.

### 3.2. Results and discussion

#### 3.2.1. Behavioral responses

We analyzed the written responses to the addition problems to determine whether the participants had interpreted the task correctly. This analysis included responses to the target (addition) trials only, because these trials are the focus of the ERP analysis. Written responses were coded as correct if both the numerical value and its semantic label denoted the sum of the two addends. For example, correct responses to the problem "*4 pears + 3 bowls* = "included "*7 objects*," "*7 things involved in eating*," and "*7 things to paint, if you're Dutch*." Responses that deviated from this standard were coded as *incorrect*. Incorrect responses included incorrect numerical responses and/or numerical responses with non-corresponding object labels. Examples of incorrect responses to the item "*4 pears + 3 bowls*" are "*7 pears*, "*7 tables*," and "*1 bowl*."

We expected some proportion of incorrect responses due to lapses of attention (e.g., misreading, miscalculation). However, a high proportion of incorrect responses could be indicative of lack of involvement with the task or of an erroneous task interpretation. Note that these two sources of error, lapses of attention and misinterpretation of the task, should be as likely to occur in aligned as in misaligned problems. A different source of errors may be the pressures associated with semantic alignment. Bassok et al. (1998) found that college students sometimes react to semantically misaligned problems by subverting the task demands. For example, when asked to produce addition word problems for functionally related object sets (e.g., *tulips* and *vases*), some participants produced division word problems instead. We reasoned that participants who understand and attend to the task might generate correct responses to semantically aligned items (e.g., *2 pies + 3 cakes* = "*5 desserts*") but generate incorrect, albeit semantically aligned, responses to misaligned items (e.g., *4 bakers + 4 pies* = "*1 baker per pie*"). Such errors, while interesting, would not be indicative of lack of engagement with the task.

Table 2 presents the percentage of correct responses to the aligned and misaligned problems for each of the 10 participants, ordered by their overall proportion of correct responses. Seven participants (numbers 1–7) show a majority of correct responses overall. Three participants (numbers 8–10) produced a lower proportion of correct responses overall but produced significantly higher proportions of correct responses for the aligned than

Table 2
Experiment 1A: Correct response rates for aligned and misaligned trials

| | | Condition | |
|---|---|---|---|
| Participant | Overall | Aligned | Misaligned |
| 1 | 0.98 | 0.98 | 0.98 |
| 2 | 0.96 | 0.98 | 0.93 |
| 3 | 0.94 | 0.98 | 0.91 |
| 4 | 0.93 | 0.93 | 0.93 |
| 5 | 0.91 | 0.89 | 0.93 |
| 6 | 0.84 | 0.73 | 0.96 |
| 7 | 0.84 | 0.93 | 0.70 |
| 8 | 0.49 | 0.93 | 0.05 |
| 9 | 0.36 | 0.68 | 0.05 |
| 10 | 0.24 | 0.41 | 0.07 |

*Note*. Participants are listed in descending order by overall correct response rate.

for the misaligned problems. This distribution suggests that all 10 participants interpreted the task correctly and were sufficiently engaged in performing it.

### 3.2.2. ERP analyses

ERPs, time-locked to the onset of each presentation of a target word, were averaged off-line for each participant at each electrode site. Grand averages were formed by averaging over participants. Trials characterized by eye blinks, excessive muscle artifact, or amplifier blocking were not included in the average. Across all participants and conditions, 4.5% of the trials were removed due to artifact.

ERP components of interest were quantified as mean voltage within four time windows: 50–150, 150–300, 300–500, and 550–800 ms. These four times windows quantify voltage for the N1, P2, N400, and P600 components, respectively. Repeated-measures analyses of variance (ANOVAS) were performed on the above dependent measures. The Greenhouse–Geisser correction for inhomogeneity of variance was applied to all repeated measures with greater than one degree of freedom in the numerator (Greenhouse & Geisser, 1959). In such cases, the corrected $p$ value is reported. Data acquired at midline, medial–lateral, and lateral–lateral sites were treated separately to allow for quantitative analysis of hemispheric differences. On the data from midline sites, two-way ANOVAS were performed, with within-subject variables of semantic alignment (aligned vs. misaligned) and electrode site. For data acquired at medial–lateral and lateral–lateral electrode sites, three-way ANOVAS were performed with within-subject variables of semantic alignment, hemisphere, and electrode site.

The waveforms observed are consistent with prior reports of ERPs to word stimuli. A clear negative–positive complex was visible in the first 300 ms following word onset (the "N1-P2" complex). These potentials were followed by a negative-going component with a peak around 400 ms (N400). We found a significant N400 effect such that target words (underlined) that were part of misaligned addition problems (*3 tulips + 5 vases*) generated larger amplitude N400s than did target words in aligned addition problems (*3 tulips + 5 roses*). Fig. 3 plots the grand-average ERPs to the target word in the semantically aligned and semantically misaligned conditions. As can be seen in Fig. 3, between 300 and 500 ms, ERPs to the misaligned targets (red line) elicited a larger amplitude N400 component than did the aligned targets (black line): midline, $F(1, 9) = 11.00$, $p < .01$; medial–lateral, $F(1, 9) = 17.86$, $p = .02$; lateral–lateral, $F(1, 9) = 11.43$, $p < .01$.

To verify that the observed differences were specific to the N400 time window, we also checked for differences between conditions in two earlier time windows (50–150 and 150–300 ms post-stimulus onset). No reliable differences between conditions were present between 50 and 150 ms. Within the 150–300 ms window, ERPs to the semantically misaligned targets were more negative-going than were ERPs to the semantically aligned targets: midline, $F(1, 9) = 16.58$, $p < .03$; medial–lateral, $F(1, 9) = 6.78$, $p < .03$; lateral–lateral, $F(1, 9) = 10.41$, $p < .02$. This difference might reflect a larger amplitude P2 in the semantically aligned condition or the onset of the subsequent N400 effect. No reliable differences between conditions were present in the P600 time window (550–800 ms).

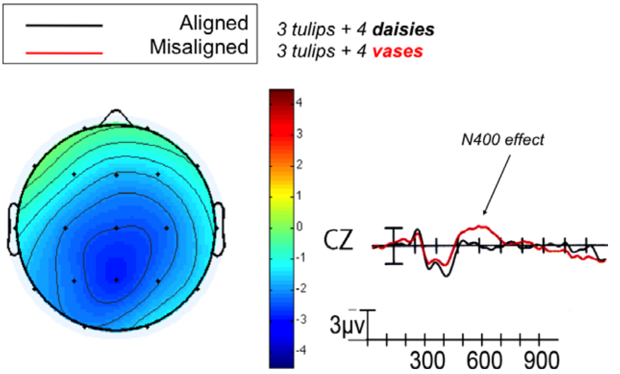


Fig. 3. Event-related brain potentials (ERPs) to target stimuli (the second object word in the sequence) in Experiment 1A, showing N400 effect triggered by semantically misaligned addition problems relative to aligned problems. Negative voltage is plotted up, the vertical calibration bar indicates target onset, and each tick mark represents 100 ms. A representative waveform is shown for the centro-posterior midline location Cz, with ERPs to aligned (black line) and misaligned (red line) target words in the context of addition problems (categorical relations for aligned, functional relations for misaligned). The topographical maps show the mean amplitude difference between the aligned and misaligned conditions in the 300–500 ms post-stimulus time window at each electrode location across the scalp (i.e., the distribution of the N400 effect).

The results of Experiment 1A confirmed our prediction that misaligned problems would yield an ERP signature distinct from that for aligned problems. In particular, we observed a greater N400 amplitude for misaligned relative to aligned problems. It appears that the addition operator coupled with the name of the first object set (e.g., *tulips*) leads to an expectancy that the second object set will be a co-hyponym of the first (e.g., roses), and that violation of this semantic constraint triggers an N400 effect. In contrast, a P600 effect was not observed.

## 4. Experiment 1B

Experiment 1A established that, in the context of addition problems, participants expect that categorically related word pairs rather than functionally related word pairs would serve as addends. This finding suggests that conceptual integration between mathematical and semantic information can occur with the speed and regularity necessary to produce the N400 effect. Taken in isolation, however, the results of Experiment 1A leave open the possibility that the observed effect was due to systematic differences in the word pairs that served as stimuli in the aligned and misaligned problems (e.g., frequency of the target words, or differences in semantic associations of the word pairs across the two conditions).

Experiment 1B, which examined conceptual integration in division problems, enabled us to address this alternative explanation. We used the same word pairs as in Experiment 1A as operand labels in division problems. Whereas addition affords relational alignment with categorically related word pairs, for division it is functionally related word pairs that align with the arithmetic relation. If any general stimulus properties (e.g., word frequency, semantic relatedness) were driving the N400 effect observed in Experiment 1A, then even in the context of division problems, functionally related word pairs should produce greater N400 amplitude than categorically related word pairs. If such an effect is not observed in division problems, then alternative explanations based on differences in general stimulus properties will be ruled out.

If a division problem privileges the expectation for relationally aligned object sets, then a problem that begins "*12 tulips ÷ 4...*," would generate a greater expectation for a functionally related term, such as *vases*, relative to a categorically related term, such as *roses*. But note that unlike addition problems, which require generation of an appropriate but unstated label for the sum (e.g., *6 tulips + 2 roses* = ***2 flowers***; *6 tulips + 2 vases* = ***8 items in a flower shop***), the *a/b* format of answers to division problems (e.g., *3 tulips/vase* or *3 tulips/rose*) allows participants to perform the task correctly using the words stated in the problem. Accordingly, the answer-generation task for division problems affords a non-modeling strategy that would yield correct answers (e.g., Martin & Bassok, 2005). Hence, there is reason to expect that the alignment effect for division problems would be smaller than the N400 effect obtained in Experiment 1A for alignment in addition problems.

### 4.1. Method

#### 4.1.1. Participants

Participants were five women and five men recruited from the University of Washington in the age range of 22–46 years (mean = 30 years). All were native English speakers with normal or corrected-to-normal vision. Participants were paid $30 for their participation in the experiment.

#### 4.1.2. Stimuli and design

Each participant responded to 141 arithmetic problems: 45 division problems with functionally related object-set labels (aligned division), 45 division problems with categorically related object-set labels (misaligned division), and 51 subtraction problems that served as fillers. The object-set labels were the same as those used in Experiment 1A.

The number pairs used in the division problems were, necessarily, different from those used in the addition problems (Experiment 1A). However, in constructing the dividend and divisor number pairs, we strove to match, as closely as possible, the variety of the numbers used in Experiment 1A and to avoid the cognitive load associated with large numbers (Ashcraft, 1992). With these goals in mind, we started with the 2 through 9 multiplication table, excluding squares (e.g., 25 ÷ 5), such that each divisor and quotient (answer) were single digits. The dividends were either one- or two-digit numbers. In order to match the relative difficulty and variety of the number pairs used in Experiment 1A, we pilot-tested candidate number pairs and selected the 21 number pairs that produced fewest errors and lowest reaction times in an answer verification task.

The same 51 filler subtraction problems from Experiment 1A were used, and the randomization and counterbalancing procedures were the same as in Experiment 1A.

#### 4.1.3. Procedure

The procedure was identical to that used in Experiment 1A, with the following exceptions: (a) Participants were told they would be solving division and subtraction problems rather than addition and subtraction problems. (b) The example problem presented to the participants was a subtraction problem (*10 animals − 2 foxes = 8 other animals*), illustrating a solution to a filler rather than a target problem. This procedural change eliminated the possibility that a worked-out solution to a division problem, with a ratio-labeled answer, might lead participants to adopt a uniform ratio-label strategy that could circumvent conceptual integration of the mathematical and semantic information. While we expected people to produce such uniformly labeled responses, to the extent possible we wanted participants to think about the appropriate answer. The ERP recording system was identical to that used in Experiment 1A.

### 4.2. Results and discussion

#### 4.2.1. Behavioral responses

As in Experiment 1A, responses were coded as correct if they were mathematically accurate and had a corresponding semantic label. Because the target trials were division problems, the numerical components of correct responses were ratios of the operands. The corresponding object labels for a quotient of two object sets could be either a ratio (*16 pears/8 bowls = "2 pears per bowl," "2 pears for every bowl"*) or a statement of relative numerosity ("*2 times as many pears as bowls*"). The latter type did not appear in the data. Responses that deviated from the above standard were coded as incorrect. Incorrect responses included those with incorrect numerical responses and/or non-corresponding object labels. For example, incorrect responses to the item "*16 pears/8 bowls = ?*" included "*2 bowls*" and "*2 fruits*." Overall, participants produced high rates of correct responses. As summarized in Table 3, those individuals who had lower rates of correct responses overall showed sensitivity to alignment conditions.

#### 4.2.2. ERP analyses

ERP data analyses were performed in the same manner as in Experiment 1A. Across all participants and conditions, 3.7% of the trials were removed due to artifact. Grand-average ERPs, averaged over all participants, revealed only small differences in the ERPs to aligned and misaligned target words (see Fig. 4). ANOVAS showed no reliable differences in the N1, P2, and N400 windows ($p > .1$ in all analyses; for N400 window, $p$ values were .24, .74., and .90 for analyses of midline, medial–lateral, and lateral–lateral positions, respectively). In the P600 window, differences between conditions approached, but did not reach, statistical significance for each of the three electrode positions ($.05 < p < .14$). However, the direction of this trend was opposite to the natural hypothesis (i.e., P600 tended to be more pronounced for aligned than misaligned condition).

Table 3
Experiment 1B: Correct response rates for aligned and misaligned trials

| Participant | Overall | Condition | |
|---|---|---|---|
| | | Aligned | Misaligned |
| 1 | 1.0 | 1.0 | 1.0 |
| 2 | 1.0 | 1.0 | 1.0 |
| 3 | 0.99 | 1.0 | 0.98 |
| 4 | 0.97 | 0.96 | 0.98 |
| 5 | 0.97 | 1.0 | 0.93 |
| 6 | 0.95 | 0.98 | 0.93 |
| 7 | 0.93 | 0.93 | 0.93 |
| 8 | 0.89 | 0.98 | 0.80 |
| 9 | 0.82 | 1.0 | 0.64 |
| 10 | 0.52 | 0.96 | 0.09 |

*Note*. Participants are listed in descending order by overall correct response rate.

Moreover, a follow-up experiment using essentially the same design and stimuli (division problems) found no evidence of either N400 or P600 differences (Guthormsen, 2007, Experiment 2B). Accordingly, the P600 trend observed in Experiment 1B appears not to be reliable.

Thus, in Experiment 1B, if participants' brains reacted differently to aligned and misaligned objects in division problems, the difference was not sufficiently uniform across trials and participants to elicit reliable N400 or P600 effects. Of course, caution is warranted in interpreting null findings. Research using behavioral tasks, such as constructing and solving word problems, shows that people do show a preferential expectation for functionally related object sets in division problems (Bassok et al., 1998; Martin & Bassok, 2005). Therefore, it is implausible to suggest that the participants in the current experiment did not have a preferential expectation for functionally related object sets given the division context.

There are several possible reasons why misalignment did not yield reliable ERP effects for division problems. Experiment 1B may have been under-powered (although the same number of participants yielded a robust N400 effect for addition in Experiment 1B). As we mentioned earlier, participants could have used the two words stated in the problem (e.g., *tulips*, *vases*) to generate correct ratio labels (*tulips/vase*), irrespective of whether or not they engaged in mathematical modeling. Another possible factor that could have contributed to the absence of an N400 effect in generating answers to division problems is the semantic specificity of the constraints on the second object label. When evaluating the alignment of an addition problem, the first object label tightly constrains the possibilities for the second object label. For example, in the case of "*10 tulips + 2* ___," an aligned label for the second object would have to be some other type of flower (e.g., *daisy*). In the case of division, the constraints on the second object label are not as strong. For example, "*10 tulips/2* ___" allows for the second object label to be a number of different types of items associated with a number of applicable relations that align well with division (e.g., *contain: vases*, *give: lovers*, *grow in: nurseries*). Therefore, for division

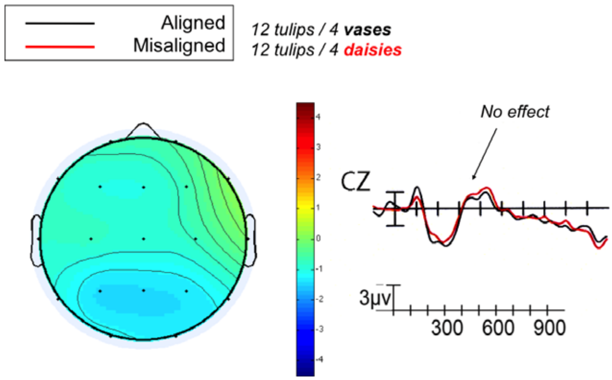


Fig. 4. Event-related brain potentials (ERPs) to target stimuli (the second object word in the sequence) in Experiment 1B, showing the absence of an N400 effect for semantically misaligned division problems relative to aligned division problems. Negative voltage is plotted up, the vertical calibration bar indicates target onset, and each tick mark represents 100 ms. A representative waveform is shown for the centro-posterior midline location Cz, with ERPs to aligned (black line) and misaligned (red line) target words in the context of division problems (functional relations for aligned, categorical relations for misaligned). The topographical maps show the mean amplitude difference between the aligned and misaligned conditions in the 300–500 ms post-stimulus time window at each electrode location across the scalp (i.e., the distribution of the N400 effect).

problems it is less likely that participants would anticipate a specific completion to the expression, thus weakening any semantic "surprise" effect associated with the N400.

Although the answer-generation task did not lead to an N400 effect for the misaligned versus aligned division problems, the results of Experiment 1B clearly show a different ERP pattern than was observed in Experiment 1A using addition, even though the specific object labels were constant across the two experiments. Thus, we can rule out the possibility that the N400 effect observed for addition was due to general stimulus properties of the object labels.

## 5. Experiment 2

In Experiments 1A and 1B, participants were required to generate answers to arithmetic problems involving numbers coupled with object labels. For addition problems (Experiment 1A), this answer-generation task led to a larger magnitude N400 response when the first and second objects were not aligned with the addition operation (i.e., were not members of the same immediate category), consistent with detection of a semantic anomaly. These results are in accord with previous findings that document the fluency and automaticity of semantic alignments for the addition operation (Bassok et al., 2008). Importantly, they provide evidence that the ERP methodology can capture fluent conceptual integration across domains. However, the answer-generation task did not reveal any semantic alignment effects for division problems (Experiment 1B), perhaps because participants could use the words provided in the problems to generate the ratio labels and hence did not need to generate a semantically meaningful label.

In Experiment 2, we introduced a different paradigm, in which generation of a semantic completion was not required. Instead, participants explicitly judged whether or not verbally stated arithmetic problems are "acceptable." This verification task does not require generation of an object name for the sum or quotient; hence, semantic retrieval will not be as central in the verification paradigm as it is in the generation task. Rather, the direct focus of the verification task is on judging acceptability. The "unacceptability" of a semantically misaligned sentence such as *Ten limes plus three bowls equals thirteen* might arise due to a perception that the two objects violate the expected analogical mapping between the roles of a categorical relation and those of the addition relation (i.e., *limes* and *bowls* are not co-hyponyms, unlike *limes* and *lemons*). A mapping violation of this sort can be construed as a structural failure (Gentner, 1983), and hence it might trigger a P600 effect. Perhaps the most similar previous finding is that of Núñez-Peña and Honrubia-Serrano (2004), who observed a P600 effect for an arithmetic series in which the final term was inconsistent with the relational pattern established by the earlier terms. An arithmetic progression might be viewed as a relational schema, such that the inconsistent final term would create a mapping violation. This interpretation suggests that unlike the answer-generation task, where an N400 was found for addition but not division, the verification task may yield comparable P600 effects (elicited by the second object word in the arithmetic problem) for both arithmetic operators.

More specifically, in Experiment 2 we measured ERPs while presenting sentence-form arithmetic problems[2] to participants that varied *both* the mathematical correctness of the problems and the alignment between the semantic and arithmetic relations (e.g., aligned: *Twelve roses divided by three vases equals four*; misaligned: *Twelve bats plus two caves equals fourteen*). Note that the answer (*four*; *fourteen*) does not contain an object label. Participants pressed a button to indicate whether or not the sentence was "acceptable." We intentionally left the instructions open as to what should determine acceptability, so that participants would be free to use whatever criterion they found to be natural. Accordingly, participants had the opportunity to either spontaneously engage in mathematical modeling by attempting to coordinate the semantic relations with the mathematical operation or to avoid mathematical modeling and simplify their task by focusing only on the mathematical elements (i.e., mathematical correctness). Based on previous studies of semantic alignment (Bassok, Wu, & Olseth, 1995; Bassok et al., 1998; Fisher & Bassok, 2009; Martin & Bassok, 2005), we expected to find individual differences in people's propensity to perform modeling.

Most obviously, a sentence could be considered "unacceptable" if the numbers yielded an arithmetic error. In this case, an N400 effect would be expected for the incorrect answer. More subtly, a sentence could be deemed structurally unacceptable if the object labels were misaligned with the arithmetic relation, as in the *bats plus caves* example above. However, semantic misalignments would only be expected to have an impact for participants who attempt to perform mathematical modeling (i.e., try to understand how the mathematical expression could make sense in a real-world situation). We predicted that those participants who engage in mathematical modeling would show P600 effects, elicited by the second object word in the arithmetic sentence due to semantic misalignment. For those participants who simply ignore the real-world meaning of mathematical expressions, alignment would not be expected to influence the ERP pattern; rather, only mathematical correctness would matter. If participants who perform mathematical modeling show a P600 effect when judging the acceptability of semantically misaligned sentences, this result would support the hypothesis that preferred semantic alignments can trigger a kind of analogical mapping.

### 5.1. Method

#### 5.1.1. Participants

The participants were 38 volunteer undergraduate students, graduate students, and staff from the University of Washington (17 female, 21 male) with a mean age of 22 years ($SD = 4.98$) who were right-handed native English speakers. Participants were either given course extra credit or paid $30 for their participation.

#### 5.1.2. Stimuli

The stimuli were simple arithmetic addition and division problem sentences that included object word pairs (e.g., *Twelve limes plus three lemons equals fifteen*) that were either categorically related or functionally related. The arithmetic problems were composed of two operands and satisfied a number of constraints required for our experimental manipulations. First, the two operands could be both added and divided to yield a whole-number answer (e.g., $12 + 3$; $12/3$). Second, we excluded tie problems (e.g., $2 + 2$) and problems containing a 1, 0, or 10 as an operand, as evidence from prior work suggests that these types of problems are processed differently, and often more easily, than other simple arithmetic problems (Ashcraft, 1992; McCloskey, 1992). Third, we only selected problems that fell into the “small” category of division problems, defined as having a divisor lesser than 25, in order to avoid issues associated with the problem-size effect (see Zbrodoff & Logan, 2005; for a review). Finally, we controlled for answer parity (LeMaire & Reder, 1999).

The object pairs were selected based on pilot testing, as described below. We initially constructed a set of 163 word pairs that we considered as belonging to one of the two semantic categories, categorical or functional, based on previous work by Bassok and colleagues (e.g., Bassok et al., 1998). The set included 83 possible categorical pairs and 80 possible functional pairs consisting of concrete, plural nouns (e.g., *cats*, *dogs*). From this set, we constructed rating surveys that were completed by 202 undergraduate students at the University of Washington as part of a class activity. Instructions asked students to rate, on a 7-point scale, either the extent to which the word pairs were categorically related or the extent to which they were functionally related, with descriptions and examples provided for each type of relation. The average categorical and functional ratings in these two conditions were compared for each word pair using an independent $t$ test with an alpha level of .05. In order to be included in the final set, word pairs had to have significantly different categorical and functional ratings, and an average rating of greater than 5 for one dimension and 4 or less for the other. Based on these ratings, we selected 48 categorical and 48 functional pairs. The word pairs in both relation conditions were equated on the average number of syllables and letters in each word.

*5.1.3. Design*

Operation (addition vs. division) was manipulated between participants ($N_{Addition}$ = 19; $N_{Division}$ = 19; participants were randomly assigned). Semantic alignment of the mathematical operation and the object sets (aligned vs. misaligned), and mathematical correctness of the problems were manipulated within participants. Correctness had three levels: correct vs. "close" incorrect vs. "other" incorrect. The "close" incorrect answer for both operations was derived by adding or subtracting 1 or 2 from the correct answer (e.g., *12 + 3 = 14*). The "other" incorrect answers for addition were the correct answers to division problems with the same operands (e.g., *12 + 3 = 4*), and the "other" incorrect answers for division were the correct answers to addition problems with the same operands (e.g., *12/3 = 15*).

For the addition problems, all of the aligned stimuli were categorically related object sets, and all of the misaligned stimuli were functionally related objects sets; the reverse was true for the division problems. Thus, the same object sets were used as argument labels for both operation conditions, but for one operation the object sets belonged to the aligned condition and for another they belonged to the misaligned condition.

The experiment consisted of three blocks of trials, with 96 trials in each block, for a total of 288 trials. Within each block, 50% of the trials were aligned word problems, and 50% were misaligned. Within each alignment type, 50% were mathematically correct, 25% were "close" incorrect, and 25% were "other" incorrect. Trial order was pseudo-randomized within each of the three blocks. Each of the paired object-set labels appeared once per block, combined with different arithmetic problems in each block.

*5.1.4. Procedure*

Participants were seated comfortably in front of an 18" CRT monitor in an isolated room and fitted with EEG recording equipment. The events on a single trial are schematized in Fig. 5. Each trial consisted of a fixation point (500 ms), and each item of the problem sentence was presented alone on a screen (450 ms with 350 ms ISI). The final inter-stimulus interval before the appearance of the YES/NO response screen was 1,000 ms (total trial duration was 7.1 s). Participants were given a hand-held controller and were asked to respond *YES* (response hand counterbalanced) using one button if they thought the problem was completely "acceptable" and *NO*, using another button if the problem was "unacceptable" in any way. They were told that the instructions were intentionally vague because the criteria by which they would judge the problems were at their discretion.

The sentences used in the acceptability task did not include object labels attached to the numerical answer (in contrast to the answer-generation task used in Experiments 1A and 1B). Participants were asked not to blink between the onset of the fixation point and

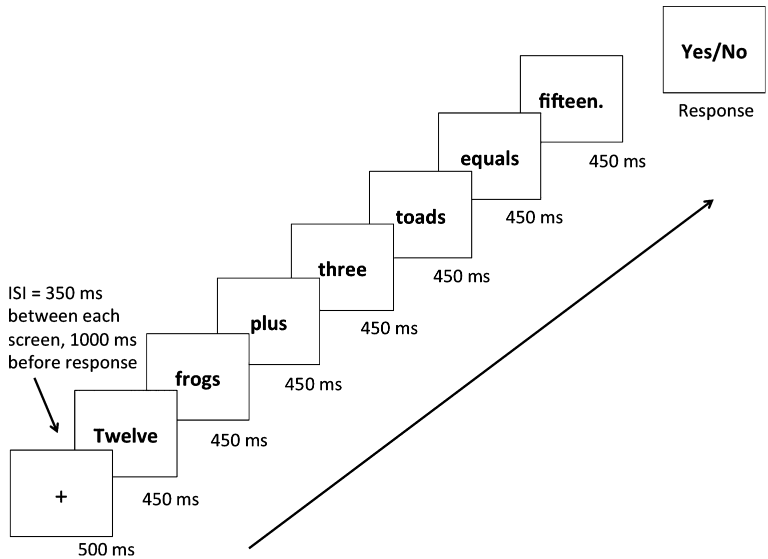


Fig. 5. The sequence of stimuli presented during trials in Experiment 2.

the appearance of the response screen. They were permitted to blink and take a short break while the response screen was displayed. Response time was not recorded, and the response triggered onset of the next trial. A break was given after each block.

### 5.1.5. EEG recording

Continuous EEG was recorded from 19 tin electrodes attached to an elastic cap (Electro-cap International) in accordance with the extended 10–20 system. Vertical eye movements and blinks were monitored by two electrodes, one placed beneath the left eye and one placed to the right of the right eye. The 19 electrodes were referenced to an electrode placed over the left mastoid, were sampled at 200 Hz, and were amplified and passed through a bandpass filter of 0.01–40 Hz (3 dB cutoff). Impedances at scalp and mastoid electrodes were held below 5 μΩ. On average across all participants, approximately 9% of trials were removed prior to averaging due to blinking and other artifacts (with exclusion of one participant; see note below). Stimuli were displayed to participants on an 18" CRT monitor approximately three feet from the participants at eye-level (when seated) with white font on a black background. The entire experiment time, including set-up, was <2 h.

## 5.2. Results and discussion

One participant (male) in the division condition was excluded from all analyses due to an extremely high artifact rejection rate (34.5% of all trials). Thus, all of the results reported below reflect the data from the remaining 37 participants (19 in the Addition condition and 18 in the Division condition).

### 5.2.1. Behavioral responses

Because the task was ambiguous in that we told participants to determine their own criteria for how to judge the problems, it is more appropriate to discuss participants' acceptability judgments response patterns with respect to our manipulations (rather than to discuss accuracy rates). Table 4 shows the percentage of problems deemed "acceptable" across the conditions for each variable (alignment, operation, and correctness). Overall, participants found aligned correct problems to be highly acceptable (mean of 90.5% of trials) and aligned problems that were mathematically incorrect (both "close" and "other" conditions) to be highly unacceptable. In the misaligned condition, the pattern is very different. Overall, misaligned correct problems were only judged to be "acceptable" on 56% of trials.

Examining the responses more closely at the participant level, participants appeared to respond either "acceptable" or "unacceptable" to misaligned correct problems in a highly reliable manner. In fact, participants can be divided into two clear groups based on their responses. Just over half of the participants ($N$ = 20, 11 from addition and 9 from division conditions) responded consistently that problems in the misaligned correct condition were "acceptable" ($M_{Acceptable}$ = 94%, $SD_{Acceptable}$ = 8%), where "consistent" was defined as responding "acceptable" on over 70% of the trials. The other participants ($N$ = 17, eight from the addition condition and nine from the division condition) responded in the opposite pattern ($M_{Acceptable}$ = 11%, $SD_{Acceptable}$ = 8%), consistently indicating that misaligned correct problems were "unacceptable." Five participants in the "unacceptable" response group for the addition condition also consistently responded that misaligned problems with the "other" mathematically incorrect answer were "acceptable." Based on informal follow-up interviews with two of the latter participants, this response pattern could possibly be due to the semantic object relations "winning out" in those participants' acceptability judgment criteria, such that the participants re-construed an addition problem as a division or "combination" problem.

In discussing the remainder of our results, we will report ERP data for these two response groups separately. Participants who consistently responded that problems in the misaligned correct condition were "acceptable" will henceforth be referred to as the "Non-modeling" group, as their acceptability judgments indicate that they did not spontaneously integrate the semantic and arithmetic relations when determining problem acceptability (e.g., *Twelve bats plus two caves equals fourteen—ACCEPTABLE*). The group

Table 4
Percentage of judgments indicating that the problem was "Acceptable," for alignment, operation, and correctness conditions in Experiment 2

| | Aligned | | | Misaligned | | |
|---|---|---|---|---|---|---|
| | Correct (%) | Close (%) | Other (%) | Correct (%) | Close (%) | Other (%) |
| Addition | 88 | 6 | 3 | 57 | 3 | 21 |
| Division | 93 | 2 | 1 | 54 | 1 | 0 |
| Total | 91 | 4 | 2 | 56 | 2 | 11 |

who consistently responded that misaligned correct problems were "unacceptable," will be called the "Modeling" group, as their responses reflected both the mathematical correctness of the problems as well as the fit between the semantic and arithmetic relations (e.g., *Twelve bats plus two caves equals fourteen—UNACCEPTABLE*). It should be noted that if data across both response groups are collapsed, effects are found that resemble that for the Modeling group, only attenuated.

### 5.2.2. *Comparison of ERPs for mathematically correct vs. incorrect problems*

Based on previous research, we hypothesized that mathematically incorrect answers should elicit an N400 response regardless of whether participants engaged in mathematical modeling. This section describes the effects found when comparing mathematically correct and incorrect problems within each participant group.

*5.2.2.1. ERP responses for Modeling group*: Each analysis described below was conducted separately at three different electrode site groupings, using electrode position and hemisphere as an additional factor in the ANOVA (midline electrodes: Fz, Cz, Pz; medial–lateral electrodes: Fp, F, C, P, O, respective to each hemisphere; and lateral–lateral electrodes: F, T, P, respective to each hemisphere).

We first compared the ERP effects elicited by mathematically correct vs. incorrect answers. Though there were some differences observed in the acceptability responses to close (e.g., *Twelve [object] plus three [object] equals* ***fourteen***) compared to other (e.g., *Twelve [object] plus three [object] equals* ***four***) incorrect answers in the misaligned condition, planned comparisons revealed no significant difference in ERP effects elicited by these two types of incorrect answers in either the aligned or misaligned condition, for either operation. Thus, these two types of incorrect answers were collapsed into an overall "incorrect" category. Fig. 6 shows ERP responses to correct and incorrect answers, plotted separately for aligned and misaligned sentences and for the Modeling and Non-modeling participant groups. As in many previous ERP sentence-processing studies, ERPs to the critical sentence-final words were superimposed over a large P300-like positive wave (Osterhout, 1997; Osterhout & Mobley, 1995). As predicted, the incorrect sums elicited an enhanced N400-like effect. This "N400 effect" actually peaked at about 300 ms, consistent with previous studies reporting the ERP response to erroneous sums in mathematical equations (cf. Jost et al., 2004).

The effect was strongest in the 250–450 ms time window, as shown in Fig. 6. All of the reported N400 results correspond to mean amplitude within this time window. The results of a 2 (correctness: correct, incorrect) $\times$ 2 (alignment: aligned, misaligned) $\times$ 2 (operation: addition, division) mixed factorial ANOVA revealed an N400 effect for mathematically incorrect problems was replicated for this group at all electrode locations [midline: $F(1, 15) = 14.68$, $MSE = 10.10$, $p = .002$; medial–lateral: $F(1, 15) = 23.82$, $MSE = 8.02$, $p < .001$; lateral–lateral: $F(1, 15) = 18.60$ $MSE = 5.97$, $p = .001$], and there was no main effect of operation at any electrode groupings. This N400 incorrectness effect appears to be followed by an LPC difference, but this is not relevant to our hypotheses and was not analyzed.

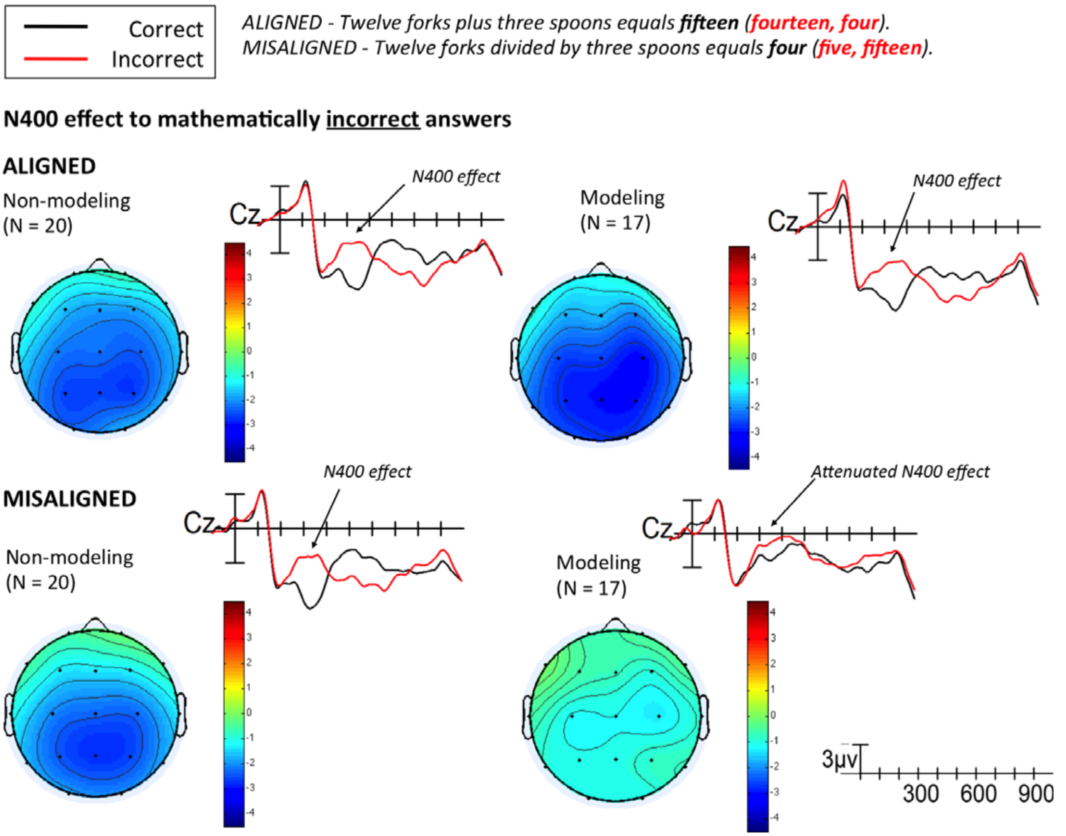


Fig. 6. Comparison of event-related brain potential effects for mathematically correct vs. incorrect problems in Experiment 2. Representative waveform graphs from the Cz electrode site illustrate the N400 effect to mathematically correct answers (black line) vs. incorrect answers (red line). The vertical calibration bar indicates target onset, each tick mark represents 100 ms, and the N400 effect is plotted separately for the aligned and misaligned conditions (manipulated within subjects) and for the Non-modeling and Modeling participant groups (determined by participant responses on the acceptability judgment task). The topographical maps show the mean amplitude difference in the 250–450 ms time window between the mathematically correct vs. incorrect answers at each electrode location across the scalp (i.e., the distribution of the N400 effect for each participant group and alignment condition). For the aligned condition, an N400 effect elicited by mathematically incorrect answers was observed in both the Non-modeling participant group and the Modeling group. For the misaligned condition, an N400 effect was observed for the Non-modeling group that was equivalent to the N400 effect observed in the aligned condition. In contrast, for the Modeling group there was an interaction between mathematical correctness and semantic alignment, such that the N400 effect elicited by mathematically incorrect answers in misaligned problems was attenuated.

Other researchers have also reported that mathematically incorrect answers elicit an N400-like negative-going wave followed by a positive-going wave (e.g., Hsu & Szucs, 2011). The proper interpretation of the positive component of the response is unclear. However, it is unlikely that the positive-going activity is related to the P600 effect elicited by syntactic anomalies, given notable differences in scalp topography. The P600 typically is largest posteriorly, whereas the math-related positivity (when it is observed) tends to be largest frontally (Hsu & Szucs, 2011).

There was also a significant main effect of alignment at all electrode groupings [midline: $F(1, 15) = 15.59$, $MSE = 33.64$, $p = .001$; medial–lateral: $F(1, 15) = 18.08$, $MSE = 55.50$, $p = .001$; lateral–lateral: $F(1, 15) = 12.64$, $MSE = 14.15$, $p = .003$], as well as a significant interaction between alignment and correctness at medial–lateral electrode sites, $F(1, 15) = 5.68$, $MSE = 14.43$, $p = .031$, and a marginally significant interaction at midline, $F(1, 15) = 3.53$, $MSE = 8.20$, $p = .08$, and lateral–lateral sites, $F(1, 15) = 4.39$, $MSE = 5.13$, $p = .054$, such that the N400 response to mathematically incorrect problems was attenuated in the misaligned condition. It is likely that for the Modeling group, the answers to mathematically correct problems in the misaligned condition were perceived as in some sense incorrect, due to the disruption in conceptual integration caused by the misaligned semantic relations encountered earlier in such problems.

*5.2.2.2. ERP responses for Non-modeling group*: In the Non-modeling group ($N = 20$), the N400 effect for incorrect answers was replicated at midline: $F(1, 18) = 22.03$, $MSE = 12.01$, $p < .001$; medial–lateral: $F(1, 18) = 27.99$, $MSE = 20.46$, $p < .001$; and lateral–lateral: $F(1, 18) = 22.71$, $MSE = 8.14$, $p < .001$ sites, and there was no main effect of operation at any electrode groupings. However, a main effect of alignment was only marginally significant at midline electrode sites, $F(1, 18) = 3.20$, $MSE = 5.84$, $p = .09$, and it was not statistically significant at the other electrode groupings. Furthermore, there was no interaction between alignment and correctness at any of the electrode groupings (see Fig. 6). The N400 effect elicited by mathematically incorrect answers, combined with the lack of interaction between correctness and alignment, suggests that participants in the Non-modeling group were actively ignoring the semantic information in the problem sentences and focusing only on the mathematical information.

### *5.2.3. Comparison of semantic information in aligned vs. misaligned problems*

Recall that we hypothesized that, due to the nature of this verification task, the second object word should elicit a P600 effect in semantically misaligned problems relative to aligned problems. However, we suspected that the existence (or magnitude) of the effect might differ depending on whether participants engaged in mathematical modeling. This section describes the effects found when comparing the object labels of semantically aligned and misaligned problems within each participant group.

*5.2.3.1. ERP responses for Modeling group*: We compared the ERP responses elicited by the second object word, which completed the semantic relation in the problem, in the aligned correct condition (e.g., *Twelve roses divided by three VASES equals four*; *Twelve limes plus three LEMONS equals fifteen*) and the misaligned correct condition (*Twelve spoons divided by three FORKS equals four; Twelve bats plus three CAVES equals fifteen*), collapsing across the addition and division conditions. We found that the second object word in the misaligned condition elicited a significant P600 effect (550–800 ms time window) relative to the aligned condition (see Fig. 7). This effect was elicited at all

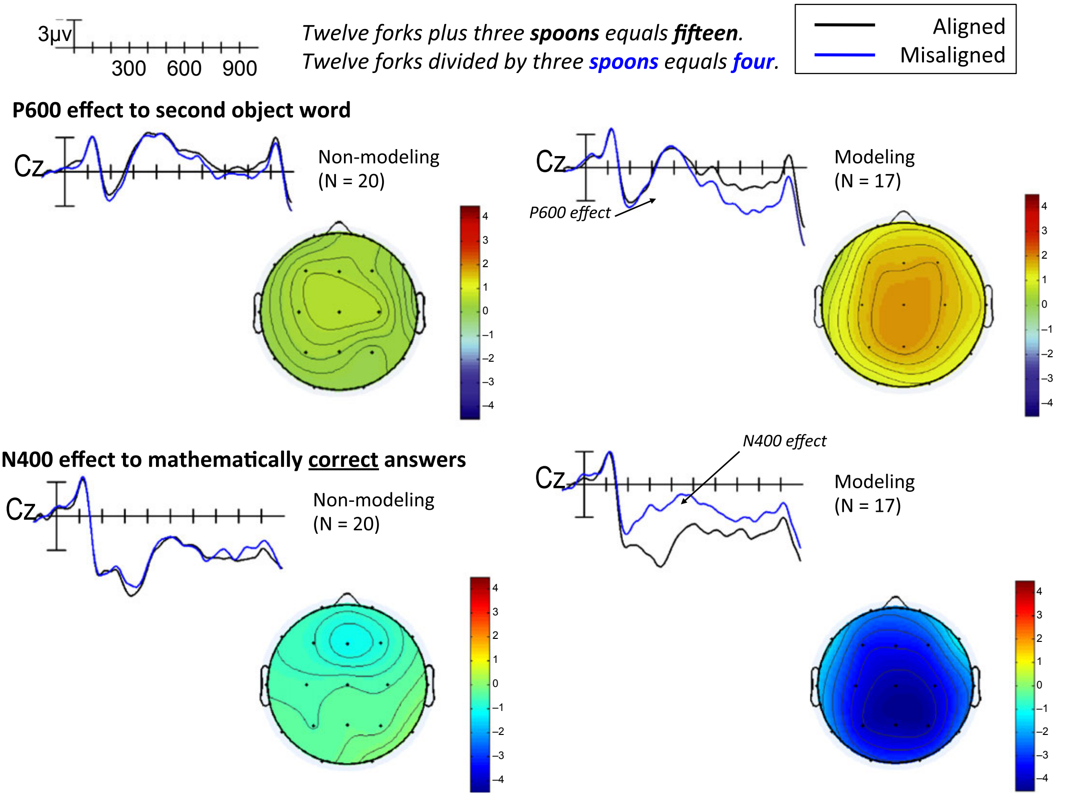


Fig. 7. Direct contrast of two different parts of aligned and misaligned problems in Experiment 2, plotted separately for each participant group (Non-modeling and Modeling). These groups were determined by participant responses on the acceptability judgment task. Representative event-related brain potential (ERP) waveforms from the Cz electrode site (the vertical calibration bar indicates target onset, and each tick mark represents 100 ms), as well as topographical distributions of each ERP effect, are shown. Topographical distribution maps show the mean difference between the aligned and misaligned conditions at each electrode location across the scalp; N400 distribution reflects the 250–450 ms post-stimulus time window and P600 distribution reflects the 550–800 ms post-stimulus time window. The top portion of the figure shows the P600 effect elicited to the second object word in semantically aligned (black line) vs. misaligned (blue line) problems. The bottom portion of the figure shows the N400 effect elicited by mathematically correct answers in the aligned condition (black line) vs. mathematically correct answers in the misaligned condition (blue line). (NB: Fig. 6 shows the effect elicited by mathematically incorrect vs. correct answers, shown separately for each alignment condition, whereas Fig. 7 figure shows the analysis that was done to directly compare alignment conditions.) The P600 (to second object word) and N400 (for mathematically correct answers) effects for misaligned relative to aligned problems were very large for the Modeling participant group, but they were not statistically reliable for the Non-modeling group.

electrode site groupings [midline: $F(1, 15) = 15.63$, $MSE = 5.18$, $p = .001$; medial–lateral: $F(1, 15) = 22.33$, $MSE = 7.05$, $p < .001$; lateral–lateral: $F(1, 15) = 22.83$, $MSE = 1.69$, $p <$ 001]. Recall that the P600 effect is typically elicited by structural violations and failures in contextual integration of thematic relations. Thus, it appears that participants in the Modeling group generally perceived the presence of misaligned semantic relations as constituting violations of the internal structure of the problem. There were no effects of operation at any of the electrode site groupings. This P600 effect is consistent with the hypothesis that analogical mapping provides a mechanism for integrating semantic and arithmetic relations.

The P600 effect (measured from onset of second object word) was accompanied by an extremely large N400 effect (measured from onset of the answer) when directly comparing mathematically *correct* answers in the misaligned condition relative to *correct* answers in the aligned condition with a 2 (alignment: aligned, misaligned) × 2 (operation: addition, division) mixed factorial ANOVA [midline: $F(1, 15) = 17.98$, $MSE = 22.24$, $p = .001$; medial–lateral: $F(1, 15) = 20.63$, $MSE = 40.21$, $p < .001$; lateral–lateral: $F(1, 15) = 13.63$, $MSE = 12.05$, $p = .002$]. While strongest in the 250–450 ms time window that corresponds to the reported ANOVA statistics, this negativity actually starts extremely early and continues throughout the waveform. There were no effects of operation at any of the electrode site groupings.

The P600 effect associated with misaligned problem sentences in the Modeling group ($N = 17$) supports the hypothesis that semantic misalignment was viewed as a structural defect in the analogical mapping between the conceptual relation and the stated arithmetic operation. The associated N400 effect for the mathematically correct answer in misaligned relative to aligned conditions indicates that the Modeling participants (those who judged misaligned problems to be "unacceptable,") also reacted to misalignment as a semantic error (similar to a mathematical error) caused by the earlier violation.

*5.2.3.2. ERP responses for Non-modeling group*: In contrast with the Modeling group, the Non-modeling group showed practically no effects of alignment. Comparing the aligned and misaligned conditions directly, there was no significant P600 effect when comparing the second object word between the aligned and misaligned conditions (see Fig. 7). However, a significant N400 effect was found comparing mathematically correct answers in the aligned condition with those in the misaligned condition over medial–lateral electrode sites, $F(1, 18) = 4.71$, $MSE = 4.91$, $p = .04$, with a marginally significant effect over lateral–lateral electrode sites, $F(1, 18) = 3.18$, $MSE = 1.87$, $p = .09$. The alignment effect for correct answers was not significant over midline sites.

## 6. General discussion

The present findings demonstrate that (contrary to a traditional view) mathematical thinking is not entirely independent of semantic knowledge. Rather, understanding of simple arithmetic statements is often based on the rapid, systematic integration of semantic relations (e.g., shared category membership of two objects) with a corresponding arithmetic operator (e.g., addition). This impact of semantic alignment on mathematical thinking has been demonstrated previously in behavioral studies, both for arithmetic operations (Bassok, 2001; Bassok et al., 1998, 2008; Martin & Bassok, 2005) and for notation types (fractions and decimals align with discrete and continuous quantities, respectively; DeWolf, Bassok, & Holyoak, 2015; Rapp, Bassok, DeWolf, & Holyoak, 2015). The present study goes beyond previous work by demonstrating, using ERP methods, that specific brain responses (N400 or P600 effects) can be triggered within a few hundred milliseconds of presentation of a word that creates a semantic misalignment. The characteristics of these brain responses, which vary with task goals and the strategic "set" of the participants, provide a more detailed understanding of the mechanisms that may support conceptual integration during mathematical thinking, notably, analogical mapping. Together, these findings support the hypothesis that conceptual integration is a domain-general cognitive process that can be used *across* semantic and arithmetic domains, and is indexed by the same ERP components that have been observed in previous studies using linguistic and other stimuli within individual domains.

### 6.1. Influence of alignment on ERP in a semantic generation task

We obtained ERP measures of semantic alignment using two types of tasks. We first used a task requiring generation of a joint semantic and numerical answer to an arithmetic problem. For example, the correct answer to an aligned problem such as "*6 pears* + *4 peaches* = "would be "***10 fruit***," whereas a correct answer to a misaligned problem such as "*4 pears* + *3 bowls* = "might be "***7 objects***." This answer-generation task thus required participants to attend to both the numerical answer and the semantics of the set generated as the output. For addition problems (Experiment 1A), a strong N400 effect to the second object label was observed in misaligned relative to aligned problems, consistent with detection of a semantic anomaly. The fact that this semantic generation task yielded an N400 effect for addition problems is consistent with the semantic basis for N400 effects observed in many language studies (e.g., Kutas & Federmeier, 2011; Kutas & Hillyard, 1980, 1984; Osterhout & Nicol, 1999). The N400 effect observed in Experiment 1A may involve a kind of relational priming (Spellman et al., 2001), where the operator (addition) coupled with the name of the first object set serves to prime the name of the second object set (expected to be a co-hyponym of the first).

Such an N400 effect was not observed for division problems (Experiment 1B), where the semantic relation between the two object sets was less constrained even in the aligned condition, and a correct answer could potentially be generated using a non-modeling strategy. Therefore, for division problems it is less likely that participants would anticipate a specific concept to fill the second mathematical role (divisor), weakening the semantic "surprise" effect associated with the N400. Thus, the presence of an N400 effect for misaligned addition problems and its apparent absence for misaligned division problems are both consistent with the interpretation that the effect is driven by the semantic constraint (or lack thereof) that the first object name in a given arithmetic operation imposes on the second. The results of Experiment 1A indicate that the principles of semantic constraint in conceptual integration apply in arithmetic word problems as well as in sentences. The failure to find semantic alignment effects in Experiment 1B (with division problems) indicates that the alignment effects observed in Experiment 1A (with addition problems) were not due to lexical-semantic associations between object labels, as exactly the same word associations were present in both experiments.

### 6.2. Influence of alignment on ERPs in a verbal verification task

In Experiment 2 we introduced a verification task to study the impact of semantic alignment. Instead of generating an answer, participants judged whether or not a statement was "acceptable." Previous studies had shown that, in arithmetic equations (e.g., *6/2 = 3*), an N400 effect is obtained for mathematically incorrect (e.g., *6/2 = 4*) vs. correct answers. Experiment 2 introduced variations in semantic alignment in sentences that expressed mathematical equations (correct or incorrect) in a purely verbal form, thereby maximizing the compatibility of the format with that used in previous studies of language processing. We measured ERPs while participants judged the acceptability of sentence-form arithmetic problems that varied *both* mathematical correctness and the alignment between semantic and arithmetic relations (e.g., aligned: *Twelve roses divided by three vases equals four*; misaligned: *Twelve bats plus two caves equals fourteen*). We intentionally left the instructions open as to what should determine "acceptability," so that participants would be free to use whatever criteria they found to be natural.

Consistent with previous studies of semantic alignment (Bassok et al., 1998; Fisher & Bassok, 2009; Martin & Bassok, 2005), as well as studies investigating mathematical modeling in other domains, such as algebra or probability (Bassok et al., 1995; Clement et al., 1981; Fisher et al., 2011), we found clear individual differences in people's propensity to perform modeling. Approximately half the participants judged misaligned but mathematically correct statements to be "acceptable" (Non-modeling group), whereas the other half judged such statements to be "unacceptable" (Modeling group). For participants in the Non-modeling group, who simply ignored the real-world meaning of mathematical expressions, alignment had no reliable influence on the ERP pattern and an N400 effect was observed for mathematically incorrect answers. In contrast, for participants in the Modeling group, misalignment elicited a P600 effect observed at the second object word that completed the semantically misaligned relation, consistent with detection of a structural anomaly. This was followed by a "last-item" N400 effect elicited by mathematically correct answers in semantically misaligned problems. Similar dual P600/N400 effects have been found in sentences that contain grammatical violations (e.g., Osterhout & Mobley, 1995).

### 6.3. Alignment as analogical mapping

The P600 effect we observed in Experiment 2 for Modeling participants making acceptability judgments is consistent with similar P600 effects observed to grammatical violations in sentences (Osterhout & Holcomb, 1992, 1995). The overall ERP pattern for these participants supports the hypothesis that structural misalignment between semantic and arithmetic relations in the word problems caused the entire problem to be perceived as incorrect. The "unacceptability" of a semantically misaligned statement (e.g., *Ten limes* *plus three bowls equals thirteen*; *Eight roses divided by four tulips equals two*) appears to arise from a perception that the two objects violate the expected analogical mapping between their roles in a semantic relation (e.g., co-hyponyms) and those of the arguments in an arithmetic operation (e.g., addition). A mapping violation of this sort can be construed as a structural failure (Gentner, 1983) of the sort that is indexed by the P600 effect, similar to a violation of a relational rule (Núñez-Peña & Honrubia-Serrano, 2004).

Given the present ERP evidence for the structural nature of conceptual integration in mathematical thinking, it is natural to hypothesize that the underlying mapping process depends on identifying consistent correspondences between objects jointly bound to roles in both semantic and mathematical relations (as depicted in Fig. 1). A number of computational models of analogical mapping have been proposed (for recent reviews see Gentner & Forbus, 2011; Holyoak, 2012). Some computational models are based on neural mechanisms (e.g., Hummel & Holyoak, 1997, 2003; Knowlton, Morrison, Hummel, & Holyoak, 2012), with particular focus on functions of the prefrontal cortex (Morrison et al., 2004). Both patient studies (e.g., Krawczyk et al., 2008; Waltz et al., 1999) and functional imaging studies (e.g., Bunge, Helskog, & Wendelken, 2009; Bunge, Wendelken, Badre, & Wagner, 2005; Cho et al., 2010; Green, Fugelsang, Kraemer, Gray, & Dunbar, 2010, 2012; Volle, Gilbert, Benoit, & Burgess, 2010; Watson & Chatterjee, 2012) have identified a network of brain areas, including subareas of the prefrontal cortex, that are active during complex relational processing.

The present findings regarding analogical alignment between semantic concepts and arithmetic operations encourage the use of ERP paradigms to examine performance in simpler analogical tasks. For example, Kmiecik and Morrison (2013) recently found an increased N400 effect for semantically distant verbal analogy problems in the proportional format. In a semantically distant problem (e.g., *father : son:: inventor : invention*), the A : B term imposes weaker semantic constraint on the C : D term than in a semantically close problem (e.g., *father : son:: mother : daughter*). Yang et al. (2013) report observing a P600 effect for incongruent metaphors. It remains to be seen whether the N400 and P600 ERP indices can be used to gain insight into the nature of the mechanisms involved in explicit analogical reasoning tasks.

### 6.4. *Individual differences in propensity to perform modeling*

The acceptability judgment task yielded clear individual differences in propensity to perform modeling. In this verification task, the participant was free to attend, or not, to the meanings of the object names. Thus, on the one hand, Non-modeling participants showed virtually no impact of alignment on their ERP pattern; on the other hand, Modeling participants showed robust ERP effects for both addition and division.

These individual differences have important theoretical implications for understanding the mechanisms underlying the influence of semantic alignment. Because ERPs are rapid—triggered within a few hundred milliseconds from the onset of a critical cue—they are usually interpreted as evidence of automatic processing (in the sense discussed by Neely, 1977). However, studies of relational priming (e.g., facilitation in reading the word pair *bear—cave* after reading *bird—nest*) suggest that such priming is not fully automatic, but rather depends on a strategic set to process relations (Spellman et al., 2001). Similarly, in tasks involving judgments of similarity, use of a mapping strategy may itself depend on the overall strategic set that the participant invokes (Markman & Gentner, 1993). The present findings are consistent with the view that semantic alignment operates as a kind of "conditional" automaticity: Given a strategic set to attend to conceptual relations (established either by the task goal or by the participant's natural propensity), ERP effects are triggered. Depending on whether the task involves generation of a semantic concept or an explicit judgment of acceptability, the form of the ERP effect (N400 or P600) triggered by semantic misalignment differs.

An important issue for further research is to determine what cognitive factors or specific mathematical abilities may predict people's tendency to spontaneously engage in mathematical modeling in situations where its use is optional. For example, consider a problem from Paige and Simon (1966): "The number of quarters a man has is seven times the number of dimes he has. The value of the dimes exceeds the value of the quarters by two dollars and fifty cents. How many has he of each coin?" Generating a mathematical equation for this problem is possible, but doing so requires problem solvers to ignore the inherent contradiction between the required operations and their real-world knowledge of American currency. Which is the "better" performance—to produce an accurate yet useless equation, or to notice that the question violates real-world knowledge? Perhaps people who have poorer mathematical abilities or lower working memory capacity may try to avoid the extra cognitive effort required by mathematical modeling, and instead simplify ambiguous, "modeling optional" tasks by just focusing on the mathematics of the problem. Alternatively, it could be the case that it is mathematical "experts" who do not spontaneously engage in mathematical modeling when they are not explicitly asked to do so, because such experts tend to think of mathematics as an abstract domain that should be independent of real-world relations. A problem solver's propensity to engage in modeling might also be dependent on whether his or her early learning environment introduced mathematical concepts by bootstrapping from semantic knowledge (e.g., using pizza slices to introduce fractions). Future research needs to more closely examine differences across individuals with respect to the propensity to perform modeling, differences across experimental contexts with respect to how misalignment is manifested, and how language comprehension and mathematical modeling interact in a broader range of situations.

## Acknowledgments

The experiments reported here encompass the PhD dissertation research of Amy M. Guthormsen (Experiment 1) and Kristie J. Fisher (Experiment 2), who are the joint first authors of the present paper. Both dissertations were completed at the University of Washington under the direction of Miriam Bassok and Lee Osterhout. Parts of the research were presented at the annual meetings of the Psychonomics Society (2008, 2009), Cognitive Neuroscience Society (2009, 2010), and Cognitive Science Society (2009, 2010). This work was partially funded by the University of Washington's Royalty Research Fund through a grant awarded to Miriam Bassok (65-3488), and by NIDCD Research Grant R01DC01947 awarded to Lee Osterhout. Portions of the paper were written while Keith Holyoak was a visiting professor at the Department of Psychology, National University of Singapore. Thanks to the members of the Cognitive Neuroscience of Language Lab for help with data collection and theoretical insights, and to Melody Sherry and Louis Wei for help with pilot data collection and analysis.

## Notes

1. The term "mathematical modeling" as used in the literature on mathematical cognition is not to be confused with its usage as a term for the development of mathematical models of empirical phenomena by theorists.
2. A preliminary experiment (Fisher, Bassok, & Osterhout, 2009) showed that arithmetic errors trigger an N400 effect when participants made acceptability judgments, both for numerical equations (as found in previous studies) and for equivalent sentences (e.g., *Three plus four equals eight*). We therefore used a sentence format in Experiment 2 to maximize consistency with the methods of previous ERP work using linguistic stimuli.

## References

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*, 75–106.

Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences, USA*, *102*(39), 14116–14121.

Bassok, M. (2001). Semantic alignments in mathematical word problems. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 401–433). Cambridge, MA: MIT Press.

Bassok, M., Chase, V., & Martin, S. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology*, *35*, 99–134.

Bassok, M., Pedigo, S. F., & Oskarsson, A. (2008). Priming addition facts with semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 343–352.

Bassok, M., Wu, L.-L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition*, *23*, 354–367.

Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage*, *46*(1), 338–342.

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex.*, *15*(3), 239–249.

Campbell, J. I. D. (1995). Architectures for numerical cognition. *Cognition*, *53*, 1–44.

Campbell, J. I. D., & Metcalfe, A. W. S. (2009). Numerical abstractness and elementary arithmetic. Commentary to Cohen, Kadosh, & Walsh (2009). *Behavioral and Brain Sciences*, *32*, 330–331.

Campbell, J. I. D., & Sacher, S. G. (2012). Semantic alignment and number comparison. *Psychological Research*, *76*, 119–128.

Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., Knowlton, B. J., & Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral Cortex*, *20*, 524–533.

Clement, J., Lochhead, J., & Monk, G. S. (1981). Translation difficulties in learning mathematics. *American Mathematical Monthly*, *88*, 285–290.

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology*, *14*(2), 218–224.

DeWolf, M., Bassok, M., & Holyoak, K. J. (2015). Conceptual structure and the procedural affordances of rational numbers: Relational reasoning with fractions and decimals. *Journal of Experimental Psychology: General*, *144*(1), 127–150. http://dx.doi.org/10.1037/xge0000034

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

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

Fisher, K. J., & Bassok, M. (2009). Analogical alignments in algebraic modeling. In B. Kokinov, K. J. Holyoak, & D. Gentner (Eds.), *Proceedings of the 2nd International Conference on Analogy* (pp. 137–144). Sofia, Bulgaria: New Bulgarian University.

Fisher, K. J., Bassok, M., & Osterhout, L. (2009). Conceptual integration in arithmetic is the same for digits and for words: It's the meaning, stupid! In N. A. Taatgen, & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2142–2147). Austin, TX: Cognitive Science Society.

Fisher, K. J., Borchert, K., & Bassok, M. (2011). Following the form: Effects of equation format on algebraic modeling. *Memory & Cognition*, *39*, 502–515.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D., & Forbus, K. (2011). Computational models of analogy. *WIREs Cognitive Science*, *2*, 266–276.

Green, A. E., Fugelsang, J. A., Kraemer, D. J. M., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*, 70–76.

Green, A. E., Fugelsang, J. A., Kraemer, D. J. M., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *38*, 264–272.

Greenhouse, S. W., & Geisser, S. (1959). On the methods in the analysis of profile data. *Psychometrika*, *24*, 95–111.

Guthormsen, A. M. (2007). Conceptual integration of mathematical and semantic knowledge. Unpublished Ph.D. dissertation, Department of Psychology, University of Washington.

Hinsley, D., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In P. A. Carpenter & M. A. Just (Eds.), *Cognitive processes in comprehension* (pp. 89–106). Hillsdale, NJ: Erlbaum.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). New York: Oxford University Press.

Hsu, Y., & Szucs, D. (2011). Arithmetic mismatch negativity and numerical magnitude processing in number matching. *BMC Neuroscience*, *12*, 83–85.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–263.

Jost, K., Henninghausen, E., & Rosler, F. (2004). Comparing arithmetic and semantic fact retrieval: Effects of problem size and sentence constraint on event-related brain potentials. *Psychophysiology*, *41*, 46–59.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*, 205–225.

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

Kmiecik, M. J., & Morrison, R. G. (2013). Semantic distance modulates the N400 event-related potential in verbal analogical reasoning. In M. Knauf, M. Pauven, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 799–804). Austin, TX: Cognitive Science Society.

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences*, *16*, 373–381.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*, 2020–2032.

Kuperberg, G. (2007). Neural mechanisms of language: Challenges to syntax. *Brain Research*, *1146*, 23–49.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*, 463–470.

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

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.

Lefevre, J., Bisanz, J., & Mrkonjic, L. (1988). Cognitive arithmetic: Evidence for obligatory activation of arithmetic facts. *Memory & Cognition*, *16*, 45–53.

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

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467.

Martin, S., & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word problem solving and equation construction. *Memory & Cognition*, *33*, 471–478.

Martín-Loeches, M., Casado, P., Ganzalo, R., De Heras, L., & Fernández-Frías, C. (2006). Brain potentials to mathematical syntax problems. *Psychophysiology*, *43*, 579–591.

McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition*, *44*, 107–157.

Mochon, D., & Sloman, S. A. (2004). Causal models frame interpretation of mathematical equations. *Psychonomic Bulletin & Review*, *11*(6), 1099–1104.

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*, 260–271.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*, 226–254.

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

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

Novick, L. R. (1990). Representational transfer in problem solving. *Psychological Science*, *1*(2), 128–132.

Núñez-Peña, M. I., & Honrubia-Serrano, M. L. (2004). P600 related to rule violation in an arithmetic task. *Cognitive Brain Research*, *18*(2), 130–141.

Nuwer, M. R., Comi, G., Emerson, R., Fuglsang-Frederiksen, A., Guerit, J.-M., & Hinrichs, H., Ikeda, A., Luccas, F. J. C., & Rappelsburger, P. (1998). IFCN standards for digital recording of clinical EEG. *Electroencephalography and Clinical Neurophysiology*, *106*(3), 259–261.

Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and Language*, *59*, 494–522.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*, 785–806.

Osterhout, L., & Holcomb, P. J. (1995). Event-related brain potentials and language comprehension. In M. D. Rugg, & M. G. H. Coles (Eds.), *Electrophysiology of mind: Event-related brain potentials and cognition* (pp. 171–225). New York: Oxford University Press.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 786–805.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*, 739–773.

Osterhout, L., & Nicol, J. (1999). On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes*, *14*, 282–317.

Paige, J. M., & Simon, H. A. (1966). Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory* (pp. 51–119). New York: Wiley.

Patel, A. D., & Daniele, J. R. (2002). An empirical comparison of rhythm in language and music. *Cognition*, *87*, 835–845.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience*, *10*, 717–733.

Pesenti, M., Thioux, M., Seron, X., & De Volder, A. (2000). Neuroanatomical substrates of Arabic number processing, numerical comparison, and simple addition: A PET study. *Journal of Cognitive Neuroscience*, *12*(3), 461–479.

Rapp, M., Bassok, M., DeWolf, M., & Holyoak, K. J. (2015). Modeling discrete and continuous entities with fractions and decimals. *Journal of Experimental Psychology: Applied*, *21*(1), 47–56. http://dx.doi.org/10.1037/xap0000036

Sitnikova, T., Holcomb, P., Kiyonaga, K., & Kuperberg, G. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience*, *20*, 2037–2057.

Sitnikova, T., Kuperberg, G., & Holcomb, P. J. (2003). Semantic integration in videos of real-world events: An electrophysiological investigation. *Psychophysiology*, *40*, 160–164.

Spellman, B. A., Holyoak, K. J., & Morrison, R. (2001). Analogical priming via semantic relations. *Memory & Cognition*, *29*, 383–393.

Szucs, D., & Csépe, V. (2004). Access to numerical information is dependent on the modality of stimulus presentation in mental addition: A combined ERP and behavioral study. *Cognitive Brain Research*, *19*, 10–27.

Szucs, D., & Csépe, V. (2005). The effect of numerical distance and stimulus probability on ERP components elicited by numerical incongruencies in mental addition. *Cognitive Brain Research*, *19*, 10–27.

Varley, R. A., Klessinger, N. J., Romanowski, C. A., & Siegal, M. (2005). Agrammatic but numerate. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(9), 3519–3524.

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene violations. *Psychological Science*, *24*, 1816–1823.

Volle, E., Gilbert, S. J., Benoit, R. G., & Burgess, P. W. (2010). Specialization of the rostral prefrontal cortex for distinct analogy processes. *Cerebral Cortex*, *20*(11), 2647–2659.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., Thomas, C. R., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*, 119–125.

Wang, Y., Kong, J., Tang, D., Zhuang, D., & Li, S. (2000). Event-related potential N270 is elicited by mental conflict processing in human brain. *Neuroscience Letters*, *293*, 17–20.

Watson, C. E., & Chatterjee, A. (2012). A bilateral frontoparietal network underlies visuospatial analogical reasoning. *NeuroImage*, *59*, 2831–2838.

West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research*, *13*, 363–375.

Yang, F. G., Bradley, K., Huq, M., Wu, D., & Krawczyk, D. C. (2013). Contextual effects on conceptual blending in metaphors: An event-related potential study. *Journal of Neurolinguistics*, *26*, 312–326.

Zago, L., Pesenti, M., Mellet, E., Crivello, F., Mazoyer, B., & Tzourio-Mazoyer, N. (2001). Neural correlates of simple and complex mental calculation. *NeuroImage*, *13*, 314–327.

Zbrodoff, N. J., & Logan, G. D. (2005). What everyone finds: The problem size effect. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 331–346). New York: Psychology Press.