Journal of Numerical Cognition
jnc.psychopen.eu | 2363-8761

Research Reports

Uncanny Sums and Products May Prompt "Wise Choices": Semantic Misalignment and Numerical Judgments

Ethan C. Brown[a], Michèle M. M. Mazzocco*[b], Luke F. Rinne[c], Noah S. Scanlon[d]

[a] Department of Educational Psychology, University of Minnesota, Minneapolis, MN, USA. **[b]** Institute of Child Development, University of Minnesota, Minneapolis, MN, USA. **[c]** School of Education, University of Delaware, Newark, DE, USA. **[d]** Department of Psychology, University of Minnesota, Minneapolis, MN, USA.

Abstract

Automatized arithmetic can interfere with numerical judgments, and semantic misalignment may diminish this interference. We gave 92 adults two numerical priming tasks that involved semantic misalignment. We found that misalignment either facilitated or reversed arithmetic interference effects, depending on misalignment type. On our number matching task, digit pairs (as primes for sums) appeared with nouns that were either categorically aligned and concrete (e.g., pigs, goats), categorically misaligned and concrete (e.g., eels, webs), or categorically misaligned concrete and intangible (e.g., goats, tactics). Next, participants were asked whether a target digit matched either member of the previously presented digit pair. Participants were slower to reject sum vs. neutral targets on aligned/concrete and misaligned/concrete trials, but unexpectedly slower to reject neutral versus sum targets on misaligned/concrete-intangible trials. Our sentence verification task also elicited unexpected facilitation effects. Participants read a cue sentence that contained two digits, then evaluated whether a subsequent target statement was true or false. When target statements included the product of the two preceding digits, this inhibited accepting correct targets and facilitated rejecting incorrect targets, although only when semantic context did not support arithmetic. These novel findings identify a potentially facilitative role of arithmetic in semantically misaligned contexts and highlight the complex role of contextual factors in numerical processing.

Keywords: priming, arithmetic, context, semantic alignment

Journal of Numerical Cognition, 2016, Vol. 2(2), 116–139, doi:10.5964/jnc.v2i2.21

Received: 2016-01-15. Accepted: 2016-04-12. Published (VoR): 2016-08-05.

*Corresponding author at: Institute of Child Development, 51 East River Parkway, Minneapolis, MN 55455, USA. E-mail: mazzocco@umn.edu

Humans are numerical thinkers. Adults often use efficient processes for generating and retrieving arithmetic sums (LeFevre, Bisanz, & Mrkonjic, 1988; Rivera, Reiss, Eckert, & Menon, 2005) that are so automatic that they can interfere with judgments on unrelated number tasks (LeFevre et al., 1988). For instance, after simultaneously viewing the cue digits *5* and *3* within a classic priming study paradigm, adults take longer to reject the target stimulus *8* as a potential match for either cue digit than to reject nearby non-sum targets, such as *4*, *7*, or *9* (Figure 1a). This "obligatory" arithmetic—the *LeFevre interference effect*—raises questions about whether adults will engage in automatic number processing when doing so is contextually inappropriate or even counterproductive. In this study, we consider how variation in semantic context interacts with the LeFevre effect, specifically whether the effect occurs only when semantic context indicates that arithmetic operations are appropriate.

Although LeFevre et al. (1988) did not consider effects of context on the interference effects they described, Bassok and colleagues did (e.g., Bassok, Pedigo, & Oskarsson, 2008). Drawing from their work on effects of

context in word problems (Bassok, Chase, & Martin, 1998) and from research on modulation of automatized cognitive processes (e.g., Besner, Stolz, & Boutilier, 1997), Bassok et al. demonstrated that the LeFevre interference effect is modulated by the nouns presented with cue digits. They proposed that categorically *aligned* noun pairs (e.g., tulips, daisies) support addition because they are appropriate to combine; whereas noun pairs do *not* support addition when they are either categorically *misaligned* (e.g., beans, planes) or related *functionally* but not categorically (e.g., pages, books). Indeed, the adults in their study showed the LeFevre interference effect *only* when cue digits were paired with *categorically aligned* nouns. The effect was absent when nouns were categorically misaligned (Figure 1b).

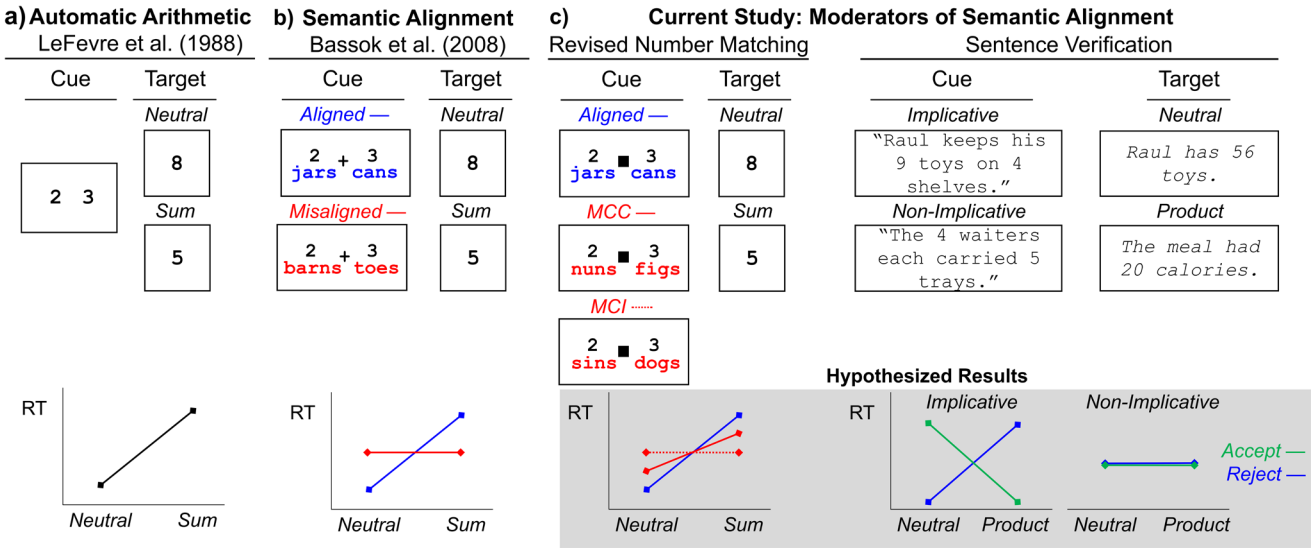

Figure 1. Priming effects for a) automatic arithmetic without context where participants match the target digit to preceding cue digits (LeFevre interference effect), b) the LeFevre interference effect moderated by the presence of categorically *aligned* or *misaligned* nouns (Bassok effect), and c) hypothesized moderators of semantic alignment examined in the present study. Our revised Number Matching task was inspired by Bassok et al. (2008); we extended it to compare types of misalignment that we differentiated as Misaligned Concrete-Concrete (MCC) and Misaligned Concrete-Intangible (MCI) noun sets. NOTE: Only rejection trials are graphed, for all three matching tasks. Our Sentence Verification task employed a similar priming paradigm, with the judgment being whether to accept or reject the target prompt statement as likely to be true, based on the cue statement, and with cue sentences either *implicating* multiplication or not.

This *Bassok effect* demonstrates that automatic arithmetic is modulated by semantic alignment. Here we modified Bassok's Number Matching task to further specify modulation of the LeFevre interference.[i] We specifically modified the misaligned condition, which in Bassok's version included multiple types of misalignment. Some of their misaligned sets included only tangible, concrete nouns (e.g., hens, radios), whereas other sets included both concrete and abstract, intangible nouns (e.g., tractors, messages). We propose that combinations of concrete and intangible referents are especially inconducive to automatic arithmetic because they are less likely to generate a plausible *rationale* for addition, compared to combinations of concrete nouns (aligned or misaligned). To test this hypothesis, we included two misaligned conditions and a categorically aligned condition in our version of the task (Figure 1c).

We also explored contextual interference with automatic arithmetic at the level of full sentences, based on evidence that semantic misalignment affects accuracy and findings of evoked related potential (ERP) responses during word problem solving and verification (e.g. Bassok, Chase, & Martin, 1998; Fisher & Bassok, 2009; Guthormsen

et al., 2016). We developed a Sentence Verification task using a priming structure similar to one used in the Number Matching task, but also built on previous semantic misalignment experiments' verification paradigms (Guthormsen et al., 2016). Our Sentence Verification task included cue sentences that either implicated multiplication (e.g. *Jill carried 2 heavy 6-packs of root beer*) or did not (e.g. *Jeff used 2 pans to make 6 omelets*), and participants judged whether a subsequent target statement was likely to be true, based on the preceding cue sentence. We tested whether the implication of multiplication in cue sentences modulated the priming effect of arithmetic products in target sentences, just as the Number Matching task tested for effects of categorical alignment on the priming effect of sums (Figure 1c). Together, these two tasks lay the groundwork for a better understanding of when adults do or do not compute numbers in context.

We also pursued two secondary aims concerning how semantic alignment effects generalize across settings and persons. First, we attempted to replicate Bassok et al.'s priming study results (2008) under conditions in which expectations for addition were less explicit, by replacing the plus sign (+) fixation point in their study with a black square. This modification provides stronger evidence that semantic context modulates obligatory arithmetic in the absence of computational symbols. Second, we explored individual differences in sensitivity to semantic misalignment and whether individuals' susceptibility to these contextual effects is associated with mathematics or reading achievement level.

Method

Participants

Participants were 92 students (61 females) enrolled in undergraduate (n = 86) or graduate programs at the University of Minnesota, who identified English as their primary language. These volunteers were predominately white (n = 64) or Asian (n = 19), and most self-reported as right-handed (n = 81). Excepting one 30 year old, the participants were 18 to 24 years old (M = 21.2, SD = 1.9). Participants opted to receive research credit or ten dollars for participating. They were naïve to the purpose of the experiment, which was described as a decision-making study. At the conclusion of the study, participants reported how many quantitative courses they had completed in college (e.g., mathematics, physics, finance) on a four-point scale. Six (7%) reported having taken no such courses, whereas others reported taking one to three (49%), four or five (24%), or six or more courses (20%).

Measures

Number Matching Task

We designed this computerized task to test whether priming for addition facts varies with semantic context, modifying the version designed by Bassok et al. (2008). Participants simultaneously viewed two cue nouns for 900 ms; immediately thereafter two cue digits appeared directly above the cue nouns for an additional 135 ms. Following these cue stimuli, two targets appeared sequentially: a single noun followed by a single digit (*noun-first* order), or vice versa (*digit-first* order). Following each target presentation, participants had two seconds to indicate via a keyboard press (Yes/No) whether the target matched either of the two preceding cue nouns or digits (see Figure 2). Noun targets were included to prevent participants from focusing solely on the numbers.

Our task differed from the original version (Bassok et al., 2008) in three ways. First, whereas Bassok and colleagues used an asterisk (*) as the initial fixation point and a plus sign (+) between the cue digits, we used a black square

(■) as both the fixation point and the symbol between cues. Second, we presented noun targets for longer durations than did Bassok et al. (900 vs. 480 ms), to reduce error rates. (In pilot studies, we confirmed that this modification reduced mean error rates from 27–30%, as observed by Bassok et al., 2008, to less than 12%.) Crucially, we presented the cue digits for 135 ms (as did Bassok et al.) in attempt to replicate the effects observed in the original study. Third, we used noun sets designed to extend Bassok et al.'s findings. As detailed further below, these included categorically aligned nouns that were appropriate to combine (such as "ticks, fleas, moths"), and two distinct types of misaligned noun sets, both of which were less appropriate to combine (such as "beans, planes, crabs"). Moreover, to rule out potential alternative explanations for priming effects, all noun sets and digit sets were constrained based on strict criteria summarized below.

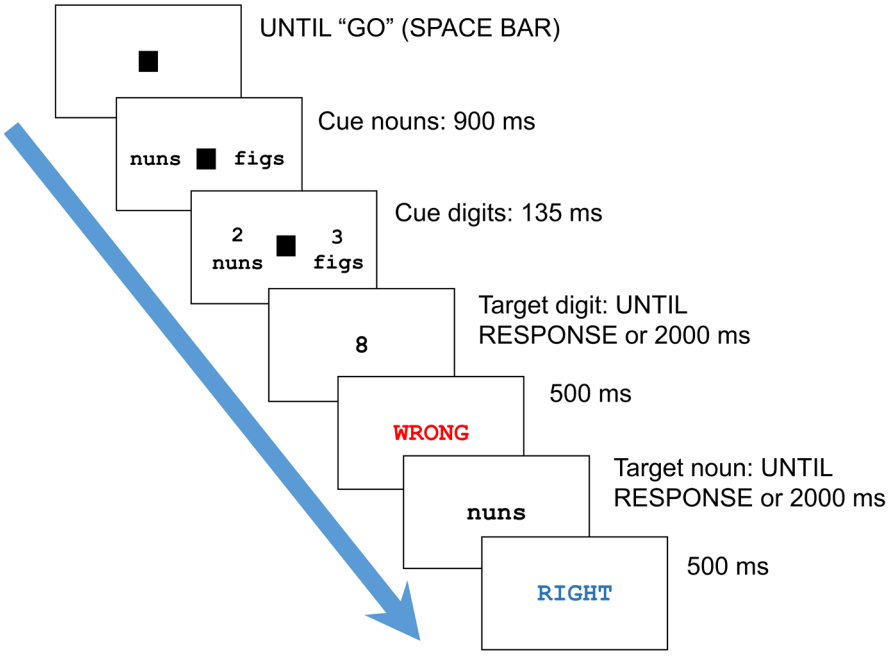

Figure 2. Illustration of the experimental procedure for the Number Matching task.

Noun sets — We included 176 noun sets and enforced strict control of the nouns' surface features, as follows: Nouns were drawn from the 5,000 most common plural nouns in the Corpus of Contemporary American English (Davies, 2010), thus excluding nouns that are rarely pluralized (e.g. throats). This list was filtered to exclude homonyms of any of the top 10,000 non-nouns. We included only enumerable nouns that were subject to regular plurals ending in 's' (e.g. we excluded irregular plurals such as "geese"). All nouns were between 4–7 letters and 1–2 syllables to limit the potential influence of noun length on participants' reaction times. Within individual noun sets, cue nouns had the same number of syllables and differed in total length by no more than one letter.

No single noun appeared in more than one noun triplet. Synonyms (e.g., ships, boats) were excluded within sets since they may be more subject to combining than other nouns. Likewise, subset relationships between cue nouns were excluded (e.g., roads, lanes) to avoid prompting division instead of addition (Bassok, Chase, & Martin, 1998). Homonyms were also excluded. To diminish familiarity effects or long-distance priming (e.g., de Vaan, Schreuder, & Baayen, 2007), each triplet appeared only once throughout the task.

Categorical alignment — Our critical manipulation was three levels of categorical alignment based on the appropriateness of summing across noun referents within a set (as per Bassok et al., 2008) and whether nouns had concrete or intangible referents (our key variable of interest). The *aligned concrete-concrete* (ACC) sets comprised concrete nouns from the same higher-level category (e.g. "orchids, poppies, lilies"). These ACC triplets correspond with the "aligned categorically (AC)" triplets in Bassok et al. (2008), and are appropriate to sum (e.g., in this case, all are flowers). Less appropriate to sum were the *misaligned concrete-concrete* (MCC) triplets, comprised of concrete nouns from different categories (e.g. "blouses, kiosks, lagoons") that refer to tangible items but otherwise differ from each other. Our third set, *misaligned concrete-intangible* (MCI) triplets, included both concrete and intangible cue nouns (e.g., raisins, chances). The MCC and MCI can be collapsed into an overall *misaligned* category that is comparable to the "misaligned unrelated (MU)" triplets in Bassok et al. (2008). For all three types of noun triplets, in half of the trials the target noun matched one of the cues, and in the other half the target matched neither cue.

The Number Matching Task included 40 noun triplets for each set (ACC, MCC, and MCI) in the digit-first trials (i.e., the trials of interest in this study). To ensure that participants did not statistically learn that noun triplets were often misaligned, 48 of the 56 noun-first trials involved ACC triplets, so that overall, half of all triplets in the study were aligned (as was the case in Bassok et al., 2008). Example nouns appear in Table 1.

Table 1

Noun Triplet Examples From the Number Matching Task

	Non-matching target			Matching target		
Noun Type	**Cue 1**	**Cue 2**	**Target**	**Cue 1**	**Cue 2**	**Target**
ACC	pigs	cows	goats	whales	sharks	sharks
	ticks	fleas	moths	doctors	lawyers	doctors
	donuts	bagels	cookies	plates	bowls	bowls
	bankers	actors	sailors	frogs	toads	toads
	apples	lemons	mangoes	lamps	desks	lamps
MCC	webs	eels	cabs	magnets	wizards	magnets
	homes	bones	cops	hotels	ladders	hotels
	cards	boats	ports	prisons	oysters	oysters
	brakes	swords	phones	statues	cigars	cigars
	robots	towels	guitars	papers	hunters	hunters
MCI	tanks	myths	laws	wives	facts	wives
	trucks	nights	clowns	pearls	tales	pearls
	sins	dogs	chips	pastors	options	options
	weeks	bombs	hairs	defects	turkeys	defects
	tactics	acorns	lessons	tasks	eggs	eggs

Note. ACC = aligned concrete-concrete; MCC = misaligned concrete-concrete; MCI = misaligned concrete-intangible.

Digit sets — Digit sets used in the Number Matching task consisted of three unique digits between 1 and 9 (Table 2). We used a subset of the digit sets created by LeFevre and Kulak (1994), which were composed of two cue digits and the target digit. Ties (e.g., "7, 7") were excluded from all cue digit pairs since these may prompt different response patterns for participants (LeFevre et al., 1988). Such restrictions limited the number of possible combi-

nations, leaving 40 distinct digit sets in the experiment. Each digit set appeared in three of the digit-first trials and either once or twice in the control word-first trials.

Table 2

Digit Triplets for Matching and Nonmatching Conditions in the Number Matching Task

Nonmatching			Matching			
	Target		Target Control		Cue Control	
Cue	Sum	Neutral	Cue	Target	Cue	Target
2■3	5	8	7■5	5	2■3	2
3■2	5	7	5■8	5	3■2	3
2■5	7	9	3■7	7	2■5	5
5■2	7	9	9■7	7	5■2	2
6■2	8	5	5■8	8	6■2	2
5■3	8	6	9■8	8	5■3	5
4■3	7	9	7■9	7	4■3	4
3■5	8	6	8■4	8	3■5	3
6■3	9	7	9■1	9	6■3	3
5■4	9	7	9■6	9	5■4	4

Note. The symbol ■ was used as a focal point for participants during this task. Each triplet comprised one cue pair, followed by a target digit.

Nonmatching digit sets. — For twenty of the 40 digit sets, the target digit did *not* match either of the two cue digits. In these nonmatching sets, the target was either a *sum* of the preceding cue digits or was *neutral* (i.e., the target did not match either of the cue digits nor their sum, product, quotient, or difference). Each of the ten pairs of cue digits was included in two nonmatching triplets, once with a sum target and once with a neutral target. The variable of interest was the difference in response latency (RT) between sum and neutral digit-first trials, a key outcome for assessing the presence of the LeFevre interference effect.

We also controlled for the size and distance between each sum/neutral target and its associated cue digits, and whether target digits appeared on the *left* or the *right*, because these factors may affect response times in ways unrelated to the LeFevre interference effect. Sum and neutral targets had a similar distribution of the digits 5 to 9 (Both: $M = 7.3$, $SD = 1.4$). The distance from the farthest cue digit to the target (the *minimum split*) was similar on average for both sets (both: $M_{neutral} = 2.8$, $M_{sum} = 2.6$), but the standard deviation for the neutral set ($SD_{neutral} = 1.7$) exceeded that of the sum set ($SD_{sum} = 1.1$), because sum splits had a unimodal distribution whereas neutral splits had a bimodal distribution. Similar patterns held for the maximum and average splits.

Matching digit sets. — Twenty of the 40 digit sets were *matching sets*, wherein the target digit matched either of the two cue digits (e.g., cue digits: 4, 6; target digit: 4). These throw-away control trials were included to verify that participants were on task and ensure that participants would not expect non-matching trials to be more likely. Moreover, to ensure that specific cue digits did not reveal whether a match was likely, ten *cue-control* triplets each had the same cue digits as one of the nonmatching sets, but also had a target digit that matched one of the cue digits.

Similarly, it was necessary to prevent the target digit from revealing whether it was likely to be a match. Since the cue digits in the nonmatching sets were constrained to sum to less than 10, these cue digits tended to be small

(M = 3.7), allowing high targets to indicate a non-match by default. Therefore, *target-control* triplets each had the same target as one of the sum triplets, but appeared following a new pair of cue digits, one of which actually matched the target. As noted above, sum and neutral targets had a similar distribution of digits, so this set also had a similar distribution of digits to neutral targets.

Test administration — Participants completed 176 trials. On the 120 *digit-first* trials relevant to this study, the digit target appeared after the cue duration and before the noun target. Thus, most of the trials appeared in digit-first order to facilitate capturing the short-term time scale of the LeFevre interference effect. In the remaining 56 control trials, the target noun appeared first (*word-first trials*) to ensure that participants attended to the words; data from these word-first trials were not analyzed.

Participants were assigned to one of four fixed trial orders. To prevent participants from ignoring word cues (since noun-first trials were less numerous), practice trials and the first of four blocks of testing trials included an equal number of digit-first and noun-first trials. However, maintaining this balance throughout the entire task would require an unfeasibly long task, so we decreased the ratio of noun-first to digit-first trials harmonically to one-half, one-third, and one-quarter in subsequent blocks for *Orders A1* and *A2*. To test whether this sequence of blocks affects participant response patterns, for *Orders B1* and *B2* the ratio of subsequent blocks was instead one-quarter, one-half, and one-third. To allow for testing of block-specific order effects, Order A2 was generated from Order A1 by switching the first 60 digit-first trials with the last 60. Order B2 was generated from Order B1 in an analogous way.

Within each order, the sequence of trials was randomly generated with several constraints. Consecutive identical answers (match vs. non-match) did not exceed four trials, and no more than four trials included the same noun- or digit-triplet type. No more than four noun-first trials occurred in a row so that these less-numerous control trials were sufficiently spread out throughout the task.

Each trial occurred in a fixed order (Figure 2). First, the fixation box appeared. Participants pressed the space bar to initiate the trial. The two noun cues appeared for 900 ms inside boxes on either side of the fixation box. Then the two digit cues appeared above the nouns, for 135 ms. Once the cues disappeared, the target (digit or noun) appeared. Participants pressed one of two color-coded keys to indicate whether the target had previously appeared as a cue (v), or had not (n). If 2 seconds passed without a response, the trial was recorded as wrong. Accuracy feedback (`RIGHT` or `WRONG`) appeared in the center of the screen for 500 ms. Then, the second target (noun or digit) was presented and participants made the same type of judgment and received feedback.

Participants received verbal instructions to strive for both accuracy and speed. They completed a demonstration trial with an experimenter who provided instructions and feedback, then completed 10 practice trials. During the experiment, participants were alerted when one-third and two-thirds of the trials were completed. Consistent with procedures adopted by Bassok et al. (2008), participants were told that they would receive a memory test; this was done to encourage attention to the nouns. The results of the memory test did not relate to the purposes of the study and were not recorded or analyzed. The task took approximately 20 minutes. All participants completed this task.

Sentence Verification Task

We created the Sentence Verification Task (Figure 3) to assess differences in priming for number-pair products as a function of whether naturalistic linguistic contexts implicate multiplication, and to test whether automaticity of number combinations (fact retrieval or rapid computation) is reduced when multiplication is clearly not implicated.

Stimuli consisted of 32 cue sentences, each followed by two prompt statements. For each trial, participants first saw a fixation box, read the cue sentence, and then saw and responded to both prompt statements sequentially, via key press (“Yes”/”No”), to indicate if the prompt was likely to be true *based on the cue sentence*. Responses to the first target prompt were analyzed; responses to the second, filler prompt were not analyzed. (See Table 3 for sample cue sentences and prompts.)

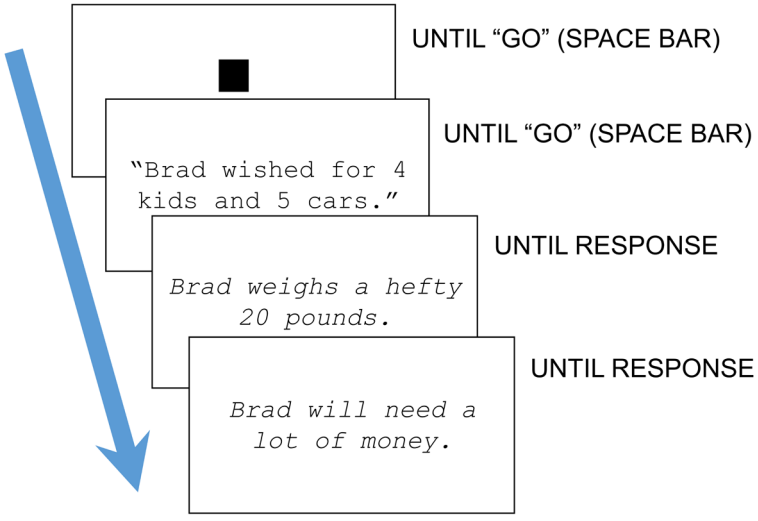

Figure 3. Illustration of the experimental procedure for the Sentence Verification Task.

Cue sentences — Each cue was a declarative sentence containing two whole numbers. The semantic content of each cue sentence either implicated multiplication of the two numbers (e.g. *Frank sent 4 texts to each of 10 friends*), or did not implicate multiplication (e.g. *You can smell those 4 pizzas 10 blocks away*). Numbers in the cue sentences ranged from 2 to 9 and were associated with multiplication facts of moderate difficulty. We excluded identical number pairs and the numbers 1 and 0 since multiplication with these numbers is relatively easy, and excluded 7 or 8 unless either was paired with 2, since products of 7 and 8 are relatively challenging (e.g., Campbell & Graham, 1985). Lengths of cue sentences were constrained in terms of number of syllables (M = 10.5, SD = 1.7, range 8–14) and words (M = 8.6, SD = 1.1, range 7–10). Numbers always appeared in Arabic notation and never began, ended, or appeared consecutively within the sentence.

Similar to the Number Matching task, the key contextual distinction between the cue sentences was whether the sentences implicated multiplication of the two numbers appearing in the sentence. In the 16 *implicative* trials, the cue sentence rendered the product of the two numbers meaningful and relevant. For instance, given the cue sentence, *The 4 waiters each carried 5 trays*, multiplying 4 by 5 yields a meaningful product. In the remaining 16 *non-implicative* trials, multiplying the numbers in a cue sentence would not yield a meaningful product (e.g., *Brad wished for 4 kids and 5 cars*).

Target prompt statements — The target prompt statements always included one number that was either the product of the two numbers from the preceding cue sentence (for 16 *product* trials) or a different number (for 16 *neutral* trials), counterbalanced across implicative/non-implicative contexts. Since our primary aim was to investigate whether modulation differed between multiplicative or non-multiplicative semantic contexts (analogous to the

Table 3

Example Stimuli for the Sentence Verification Task

Prompt Type	Digit Type	Cue Sentence	Target Prompt	Filler (Second) Prompt
Implicative Contexts				
Reject	Neutral	"The 3 ships each transported 9 crates."	*The ships weighed 5 ounces.*	*The crates were made of silk.*
Reject	Product	"Evan mowed 10 lawns at 6 dollars each."	*Evan mowed 60 lawns.*	*Evan was 6 years old.*
Accept	Neutral	"Frank dealt 4 cards each to 10 poker players."	*The full deck had 52 cards.*	*The cards were on fire.*
Accept	Product	"Gwen bought 6 toys for each of her 4 kids."	*She purchased 24 items.*	*Gwen's kids were goats.*
Non-Implicative Contexts				
Reject	Neutral	"Jacky visited 3 orchards to pick 9 peaches."	*Each peach had a 4 lb pit.*	*Jacky had enough peaches for 2 pies.*
Reject	Product	"They reserved 6 tables at the 4 Seasons Cafe."	*They reserved the tables for 24 days.*	*The reservations were for a golf tee time.*
Accept	Neutral	"Don baked 3 batches of lemon squares in 9 pans."	*Each lemon square had 4 corners.*	*Don used sugar in his baking.*
Accept	Product	"Mike won 6 medals in 4 hours."	*There were 24 hours in each day.*	*The medals he won were invisible.*

Note. Implicative trials implied a multiplication operation on the two numbers in the cue sentence. *Reject* trials had target prompts that were keyed as false and *Accept* trials were keyed as true. *Product* trials had target prompts that contained the product of the two numbers in the cue sentence; *Neutral* trials did not. Responses to the second prompt were not analyzed.

Number Matching task), we expected products to facilitate accepting, or interfere with rejecting, the prompt, as an extension of the LeFevre interference effect. We made target prompts slightly shorter than the cue sentences, in terms of both syllables (M = 8.0, SD = 1.5, range 5–11) and words (M = 5.8, SD = 1.3, range 4–10). The number embedded in the target prompt was an Arabic numeral between 2 and 60, and it never appeared at the beginning or end of the sentence.

One half of target prompts were classified as *accept,* and the other half were classified as *reject*. For instance, following the cue sentence, "*The 4 waiters each carried 5 trays*," the prompt, "*The meal had 20 calories*," was designed to be rejected, as it does not follow from the cue sentence. (Indeed, all pilot participants rejected this statement.) Including both classifications (accept and reject) across conditions ensured that participants could not statistically infer that either response was more frequent and also allowed us to examine possible differences in priming between acceptance and rejection responses. Our classifications were validated during pilot testing, and only items with greater than 80% accuracy among pilot participants were retained.

We analyzed responses for the first prompt sentence only, because priming effects on the second prompt sentence may have been contaminated by the presence of the first prompt sentence. The second prompts included both *accept* (50%) and *reject* (50%) trials; the trials either did (10 of 32) or did not (22) contain a number, to prevent participants from anticipating numbers in all prompts or in only the first prompt.

Experimental trials — Eight sets of cue sentences and target prompt statements were generated in a 2 (Prompt Type: Accept or Reject) × 2 (Context for Products: Implicative or Non-implicative) × 2 (Digit Type: Neutral or Product) design. There were four sentences per experimental condition, yielding a total of 32 trials.

Several features of the sentences were balanced across conditions in order to strengthen the validity of reaction time (RT) comparisons. Each experimental condition had exactly one trial in which the first prompt referred to the same unit of measurement as the cue. For instance, there were four trials wherein cues implicated multiplication but the product did not appear in the first target prompt (e.g., the cue sentence, "Frank dealt 4 *cards* each to 10 poker players" was followed by the target prompt "The full deck had 52 *cards,*" emphasis added). The three other

trials did not share units of measurement across the cue and first prompt. In each set of four product trials, the cue sentences presented the digit pairs 3 and 9, 4 and 10, 8 and 2, and 9 and 4, respectively. A different set of digit pairs was used for neutral trials: 2 and 6, 4 and 5, 6 and 4, and 10 and 6. This ensured a balance of stimuli across implicative/non-implicative and accept/reject trials.

Pilot testing — Cue sentences and prompt sentences were finalized through iterative piloting with 210 adults who completed prior versions of the task, either as volunteer study participants at our university (82 participants) or on Mechanical Turk (127 participants), an online marketplace for contract work where participants were paid for their responses. Based on pilot responses, we modified statements to maximize ease of judging the likelihood of being true. In the final pilot testing, two items were excluded for failing to reach our threshold of 80% accuracy, including an Accept/Implicative condition item (71%) and an Accept/Non-implicative condition item (59%). These were omitted because unusually difficult items may introduce cognitive complexity and construct-irrelevant variance to the measures. All remaining items had accuracy rates of 85% or above.

Administration — Participants listened to instructions, completed a single demonstration practice trial that did not involve any numbers, and then received feedback. All participants saw the same stimuli in the same quasi-randomly generated order adjusted to limit the number of consecutive trials with the same combination of condition and outcome (no more than two in a row) or the same keyed response (no more than three in a row for the first prompt). No feedback was given on trial responses. The task required about 5 minutes to complete. Two participants were excluded from analyses for failing to respond correctly to any trials in one or more conditions.

Achievement Measures

Math fluency — Participants completed a three-minute calculation fluency measure, the Math Fluency subtest of the Woodcock-Johnson III, during the testing session. This subtest is from a standardized, paper-and-pencil mathematics achievement measure. Participants were asked to solve as many problems as quickly as possible. Problems appeared in a test booklet, in order of increasing difficulty. The subtest has a median reliability of .92 with adult participants (Mather & Woodcock, 2001). Accuracy (number correct) and total time to complete the task were recorded. We calculated participants' fluency rate (trials/minute) to create a comparable measure for all participants. One participant's score was omitted due to experimenter error.

College entrance exam scores (ACT/SAT) — Participants were asked to provide their standardized college entrance examination test scores (ACT Math and ACT Reading). The ACT and SAT are widely used standardized college entrance exams. Each exam yields separate Mathematics and Reading scores. Both tests require basic to complex mathematics problem-solving skills or reading skills that tap meaning comprehension. Historically, scores for these exams have been highly correlated, with reported correlations of .92 for composite scores, .89 for Mathematics, and .83 for ACT Reading with SAT Verbal (now labeled Critical Reading; Dorans, Lyu, Pommerich, & Houston, 1997). Accordingly, we collapsed percentile score data across ACT or SAT Mathematics, and across ACT Reading and SAT Critical Reading scores. Of 73 participants who consented to our accessing their standardized test scores, 67 had taken the ACT. Therefore, the ACT scores were the focus of analysis, and 6 sets of SAT scores were transformed to align with the ACT metric using published national percentile norms. Participants took the ACT between 2006 and 2013, but ACT scale scores are constructed to be comparable across these years and can be analyzed directly (ACT, Inc., 2014).

Procedures

The study was approved by our institutional human subjects review board. All 92 participants completed the Number Matching, Sentence Verification, and Math Fluency tasks, in that order. (A matching task excluded from the present study was administered as the third of four tasks.) The entire testing session took approximately one hour. In addition, Math and Reading ACT and SAT scores were collected from 73 participants who consented for the University's Office of Institutional Research to release these scores to the researchers.

Results

We carried out separate analyses for our two primary numerical tasks. We used repeated measures ANOVAs to test for hypothesized main effects and interactions involving noun alignment in the Number Matching task (non-matching trials only) and implication of multiplication in the Sentence Verification task (responses to first prompt sentences only). For the Number Matching task, we first evaluated whether we replicated Bassok and colleagues' (2008) findings, and then tested our hypotheses concerning further contextual influences of misalignment on classic priming effects. Finally, for both numerical tasks, we used linear mixed models to test for individual differences and the contributions of math fluency and ACT scores to speed of response. In all analyses, the outcome variable of interest was speed, measured by the inverse response time, consistent with prior recommendations for RT modeling (e.g., Baayen & Milin, 2010; Ratcliff, 1993). This transformation increased the normality of the dependent variable in our Number Matching Task (skew = 0.1, kurtosis = 0.1) compared to both the untransformed data (skew = 1.5; kurtosis = 3.1) and a logarithmic transformation (skew = 0.7; kurtosis = 0.5). In the presentation of results, estimated parameters are back-transformed to reaction times (ms per trial), where possible, to aid interpretation and support comparisons to prior results in the research literature. Generalized eta-squared estimates are reported for all ANOVAs due to their comparability as effect sizes across research designs (Olejnik & Algina, 2003). We present the R^2_{GLMM} defined by Nakagawa and Schielzeth (2013) and generalized by Johnson (2014) as a measure of overall model fit for linear mixed models. We present the marginal R^2_{GLMM} to evaluate changes in fixed effects and the conditional R^2_{GLMM} to evaluate changes in random effects. These measures are not necessarily comparable to the R^2 used in linear regression and should be interpreted with caution.

Number Matching Task

ANOVAs

We first examined the degree to which our results replicate those of Bassok et al. (2008). We collapsed our two misaligned conditions to approximate the misaligned unrelated (MU) condition used by Bassok and colleagues, the latter of which included noun triplets similar to our MCC (e.g., "coats, biscuits, islands") and MCI conditions (e.g., "tractors, messages, fairies"). We then carried out a 2 (Context for Sums: Aligned vs. Misaligned) × 2 (Digit Type: Sum vs. Neutral) repeated measures ANOVA on the inverse response time (trials/s), or *speed*, on all correctly answered trials (Table 4). Digit Type referred to whether target numbers appearing after cue digits were the *sum* of the preceding cue digits or were *neutral* (matching neither cue digit nor the digits' sum, product, quotient, or difference).

Our replication attempt was successful. We found a Context × Digit Type interaction (η^2 = .005) similar in strength to that found by Bassok and colleagues (2008; η^2 = .008). Non-matching Aligned *Neutral* targets were rejected significantly faster than Aligned *Sum* targets, ΔM = 41 ms, $t(91)$ = 5.41, d = .565, Holm-adjusted p < .001, but

Table 4

Repeated Measures Analysis of Variance on Inverse Reaction Times (trials/s) for the Number Matching Task

Effect	*df*(n)	*df*(d)	*F*	*p*	η^2
Combined Misaligned Analysis					
Digit Type	1	91	24.83	<.001	.008
Context for Sums	1	91	2.25	.138	.001
Digit Type × Context for Sums	1	91	15.49	<.001	.005
Expanded Misaligned Analysis					
Digit Type	1	91	14.26	<.001	.004
Context for Sums	2	182	1.21	>.250	.001
Digit Type × Context for Sums	2	182	14.79	<.001	.010

Note. The Combined Misaligned Analysis had two levels of Context: Aligned and Misaligned. The Expanded Misaligned Analysis of the same data differentiated the Misaligned condition further by separating Misaligned Concrete-Concrete (MCC) and Misaligned Concrete-Intangible (MCI) conditions.

there was no difference in speed of rejection of non-matching Neutral and Sum targets on *Misaligned* trials, Δ*M* = 5 ms, *t*(91) = 0.89, Holm-adjusted *p* > .250, *d* = .093. Means for the Neutral and Sum trials under misaligned conditions (both ≈ 760 ms) fell between those for the Aligned Sum (*M* = 788 ms) and Aligned Neutral conditions (*M* = 747 ms), similar to the results reported by Bassok and colleagues. Presumably due to longer presentation durations for noun cues in our study, our accuracy rates for each condition (*M*s ≈ 90%) were substantially higher than those found by Bassok and colleagues (*M*s ≈ 70%), but response speeds followed the same qualitative pattern (see Figure 4a).

The findings were only partially similar when we separated the two types of misalignment (Figure 4b and Table 4). A 3 (Context for Sums: ACC, MCC, or MCI) × 2 (Digit Type: Sum vs. Neutral) repeated-measures ANOVA on correct trial speeds revealed a stronger Digit Type × Context interaction, which accounted for a greater proportion of the variance (η^2 = .010) than in the collapsed analysis (η^2 = .005). Moreover, the MCC trials displayed the classic LeFevre *interference* effect (LeFevre et al., 1988), wherein rejection of non-matching Sum targets was slower than rejection of non-matching Neutral targets, Δ*M* = 41 ms, *t*(91) = 3.22, Holm-adjusted *p* = .002, *d* = .335. This effect is substantial but smaller than that observed in Aligned trials, *d* = .565. The pattern of means for Sum trials is consistent with the hypothesis that increasing contextual support for summation leads to increasing interference (slower rejection) on Sum trials. A linear contrast testing this trend (MCI < MCC < ACC) was significant, *t*(182) = 4.81, Holm-adjusted *p* < .001, $r_{contrast}$ = .336.

However, unlike the MCC condition, the MCI condition had a *facilitative* effect, with *faster* rejection for non-matching Sum versus non-matching Neutral targets, Δ*M* = 18 ms, *t*(91) = 2.41, Holm-adjusted *p* = .018, *d* = .244, although notably weaker than the interference effects observed on ACC and MCC trials. Moreover, speeds were *slower* for Neutral MCI trials compared to both Neutral MCC trials, Δ*M* = 24 ms, *t*(91) = 3.12, Holm-adjusted *p* = .002, *d* = 0.323, and Neutral ACC trials, Δ*M* = 24 ms, *t*(91) = 3.13, Holm-adjusted *p* = .002, *d* = .326, which did not differ significantly from one another, Δ*M* = 0 ms, *t*(91) = 0.03, Holm-adjusted *p* > .250, *d* = .004.

We examined patterns of individual responses to rule out the potential influence of outliers on the observed facilitative effect of the MCI trials (under the Sum condition). Distributions and SDs were similar across conditions, and inspection of participant-level distributions of interference (Neutral – Sum) did not reveal outliers. Moreover,

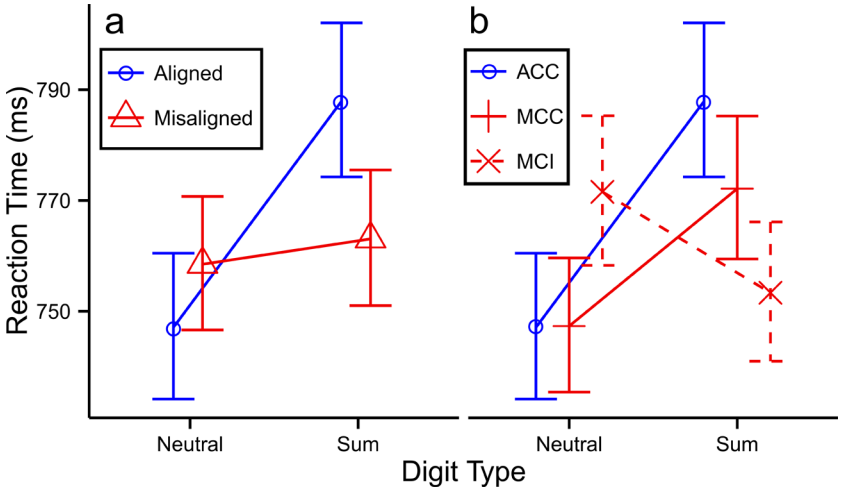

Figure 4. Reaction times in the Number Matching task (non-matching trials only) with Context for sums separated (a) two ways into Aligned Concrete-Concrete (ACC) vs. Misaligned, and (b) with Misaligned further separated into Misaligned Concrete-Concrete (MCC) and Misaligned Concrete-Intangible (MCI). Error bars represent one standard error of the mean.

a binomial sign test revealed that a statistically significant number of participants (60 of 92) displayed LeFevre interference for ACC trials, $p = .004$. Separate sign tests revealed the same result for MCC trials (60 of 92 participants), $p = .004$, and a marginal non-significant facilitative effect for MCI trials (55 of 92 participants), $p = .08$. Thus, despite lower power, the results of non-parametric binomial tests converge with the ANOVA results reported earlier.

Linear Mixed Models

Linear mixed modeling can provide additional insight into individual differences and help bring more features of the design under statistical control (e.g., Baayen, Davidson, & Bates, 2008; Bryk & Raudenbush, 1992). However, as no significant evidence emerged regarding individual differences in contextual sensitivity or our covariates, we only briefly summarize the results here and in Table 5. Table 5 shows the series of models that included the 73 participants for whom we had complete data on all covariates. (Models without covariates that included the full sample did not differ appreciably from those for the restricted sample, and are thus not reported.) Model 1 estimated a 3 (Context for Sums: ACC, MCC, or MCI) × 2 (Digit Type: Sum vs. Neutral) linear mixed model. Including random intercepts for each participant dramatically improved model fit based on a likelihood ratio (LR) test, $\chi^2(1) = 1517$, $p < .001$, Δ conditional $R^2_{GLMM} = .360$, as did including random intercepts for each item, $\chi^2(1) = 58.4$, $p < .001$, Δ conditional $R^2_{GLMM} = .021$. As with the repeated measures ANOVA, there was a significant Context × Digit Type interaction, Kenward-Roger $F(2, 54) = 3.49$, $p = .037$, Δ marginal $R^2_{GLMM} = .004$. Model 2 controls for practice and/or fatigue effects by additionally including a fixed effect for the trial number. For each successive trial, participants performed about 0.0014 trials/s faster, Kenward-Roger $t(151) = 15.89$, $p < .001$.

Math fluency and ACT scores may also capture individual differences relevant to the Number Matching task. Math Fluency had a higher correlation with response speed ($r = .19$) than did ACT Math ($r = .16$), so Math Fluency was entered into the regression first in Model 3 (results were comparable regardless of order). Each additional problem correct per minute in the Fluency measure corresponded to a 0.006 trials/s increase in speed, which only approached significance, Kenward-Roger $t(81) = 1.88$, $p = .063$. If variation in contextual sensitivity is attributable

Table 5

Linear Mixed Effects Regression Weights of Inverse RT (trials/s) on the Number Matching task (n = 73)

	Model							
	(1)		(2)		(3)		(4)	
Effect	**β**	***SE***	**β**	***SE***	**β**	***SE***	**β**	***SE***
Trial number			0.001***	0.0001	0.001***	0.0001	0.001***	0.0001
Digit Type (Sum)	–0.065**	0.026	–0.064***	0.020	–0.064***	0.020	–0.064***	0.020
Context (MCC)	0.007	0.025	–0.002	0.020	–0.002	0.020	–0.002	0.020
Context (MCI)	–0.045*	0.026	–0.036*	0.020	–0.036*	0.020	–0.036*	0.020
ACT Reading[a]							0.016***	0.005
Math Fluency rate					0.004*	0.002	0.003	0.002
Digit Type (Sum) × Context (MCC)	0.011	0.036	0.021	0.028	0.021	0.028	0.021	0.028
Digit Type (Sum) × Context (MCI)	0.092**	0.036	0.068**	0.028	0.068**	0.028	0.068**	0.028
Constant	1.300***	0.029	1.200***	0.029	1.000***	0.100	0.600***	0.160
Akaike Inf. Crit.	879		643		642		634	
Bayesian Inf. Crit.	935		706		711		710	
Marginal R^2_{GLMM}	.006		.050		.067		.109	
Conditional R^2_{GLMM}	.385		.419		.419		.417	

Note. All variables are uncentered. MCC: Misaligned Concrete-Concrete, MCI: Misaligned Concrete-Intangible. All models contained crossed random intercepts for subjects and items. Reference levels were *Aligned Concrete-Concrete (ACC)* for Context, and *Neutral* for Digit Type. All *p* values are based on *t*-tests using the Kenward-Roger approximation value for the degrees of freedom. Marginal R^2_{GLMM} estimates the variance accounted for by fixed effects while conditional R^2_{GLMM} estimates the variance accounted for by both fixed and random effects.

[a]Includes 6 participants with missing ACT Scores imputed from SAT scores.

*p < .1. **p < .05. ***p < .01.

to math fluency, we would expect to see significant interactions with condition variables; however, no interactions with Fluency approached significance.

Achievement measures were then entered into the model. ACT Math was not a significant predictor of speed, β = –0.003 trials/s, Kenward-Roger $t(80) = –0.41$, $p > .250$, but there was a significant positive main effect of ACT Reading, β = 0.016 trials/s, Kenward-Roger $t(79) = 3.21$, $p = .002$. With ACT Reading included in the model, Fluency was no longer a significant predictor, β = 0.003 trials/s, Kenward-Roger $t(79) = 1.39$, $p = .168$. Again, no interactions emerged as significant. There was no evidence of remaining unexplained individual variability in reaction times across experimental conditions, as tests of random slopes for Context, Digit Type, and their interactions were not significant in likelihood ratio tests, ps > .250. Therefore Model 4, with ACT Reading added as a predictor, was considered the final model.

In summary, we found striking evidence of interactions on the Number Matching task, including interference effects consistent with LeFevre et al. (1988) for Sum trials (in ACC and MCC conditions), modulation effects of context like those reported by Bassok et al. (2008), and an unanticipated facilitation effect on rejecting non-matching cues on MCI trials. Linear mixed models did not reveal contributions of math achievement level, but did show minor contributions of ACT Reading, despite the fact that the contextual variation in the Number Matching task was relatively artificial and did not impose significant comprehension demands. In contrast, the Sentence Verification task described next involved more authentic linguistic contexts.

Sentence Verification Task

In this task, cue numbers were presented within complete sentences that either implicated or did not implicate multiplication, and the prompt statements that followed contained either the product of the cue numbers or a neutral number. Two participants were excluded from these analyses for failing to respond correctly to any trials in one or more conditions.

ANOVAs

We first carried out ANOVAs to test whether contextual alignment moderated evaluation of the veracity of cue sentences. This 2 (Prompt Type: Accept or Reject) × 2 (Context for Products: Implicative or Non-implicative) × 2 (Digit Type: Neutral or Product) repeated measures ANOVA focused on participants' mean response speed on correct trials only. We found a strong Prompt Type × Context × Digit Type interaction, $F(1, 89) = 97.59$, $p < .001$, $\eta^2 = .053$, and significant main effects and two-way interactions, excepting the Context × Digit Type interaction (Table 6).

To further understand this three-way interaction, we evaluated Implicative and Non-implicative trials separately (Table 6). The 2 (Prompt Type: Accept or Reject) × 2 (Digit Type: Neutral or Product) repeated measures ANOVA on implicative trials showed evidence of a Prompt Type × Digit Type interaction (Figure 5a), $F(1, 89) = 7.01$, $p = .010$, $\eta^2 = .008$. Unlike the interference effect seen in Number Matching task, there was little evidence of interference with rejection of incorrect prompts for Implicative Product trials, $\Delta M = –61$ ms, $t(89) = –1.43$, Holm-adjusted $p = .156$, $d = –.151$. Participants were, however, faster to accept correct prompts on Product versus Neutral trials, $\Delta M = 311$ ms, $t(89) = 4.00$, Holm-adjusted $p < .001$, $d = .421$. This *facilitation* effect on *accepting* correct prompts on Implicative trials parallels the *interference* effect we observed for *rejecting* non-matches in the Number Matching task.

The repeated measures ANOVA on trials with Non-Implicative contexts revealed clear evidence of a strong Prompt Type × Digit Type crossover interaction, $F(1, 89) = 183.3$, $p < .001$, $\eta^2 = .145$ (Figure 5b). Participants were faster to accept correct prompts on Product trials than on Neutral trials, $\Delta M = 420$ ms, $t(89) = 5.91$, Holm-adjusted $p < .001$, $d = .623$, and were faster to reject incorrect prompts on Product trials compared to Neutral trials, $\Delta M = 557$ ms, $t(89) = 13.5$, Holm-adjusted $p < .001$, $d = 1.422$. For product prompts in the Non-Implicative condition, facilitation of correct rejection coupled with interference with correct acceptance is analogous to the effect of the MCI condition seen in the Number Matching task, where facilitation was observed for the rejection of non-matching targets on sum trials, and interference was observed for the rejection of non-matching targets on neutral trials. The effect sizes in the Non-Implicative condition for the Prompt Type × Digit Type interaction and associated *post hoc* tests were much larger than those in either Implicative trials or the Number Matching task.

Since exactly one trial in each condition had the same unit paired with one of the cue numbers and also the target number (*matched* trials), and the other three trials per condition had no units that matched in both the cue and target (*unmatched* trials), we further investigated whether these unit conditions appeared to change the results on non-implicative trials. Using a 2 (Prompt Type: Accept or Reject) × 2 (Digit Type: Neutral or Product) × 2 (Unit: Matched or Unmatched) repeated measures ANOVA on the 48 participants who had data in all cells, we found the same Prompt Type × Digit Type interaction, $F(1, 47) = 28.62$, $p < .001$, $\eta^2 = .037$, but there was no Prompt Type × Digit Type × Unit interaction, $F(1, 47) < 0.001$, $p > .250$, $\eta^2 < .001$. This indicates that the interaction is not simply due to the units associated with the digits. Post-hoc comparisons on *matched* non-implicative only trials produced qualitative patterns and large effect sizes similar to those for all non-implicative trials described above,

Table 6

Repeated Measures Analysis of Variance of the Effects of Different Contextual Stimuli in the Sentence Verification Task

Effect	*F*(1, 89)	*p*	η^2
Full Analysis			
Context	42.06	<.001	.020
Prompt Type	25.03	<.001	.031
Digit Type	32.81	<.001	.017
Context × Prompt Type	7.70	.007	.004
Context × Digit Type	1.07	>.250	.001
Prompt Type × Digit Type	35.41	<.001	.018
Context × Prompt × Digit Type	97.59	<.001	.053
Implicative Trials Analysis			
Prompt Type	5.98	.016	.011
Digit Type	16.88	<.001	.021
Prompt Type × Digit Type	7.01	.010	.008
Non-Implicative Trials Analysis			
Prompt Type	41.13	<.001	.067
Digit Type	12.47	.001	.013
Prompt Type × Digit Type	183.34	<.001	.145

Note. Full Analysis included both Implicative and Non-Implicative Trials.

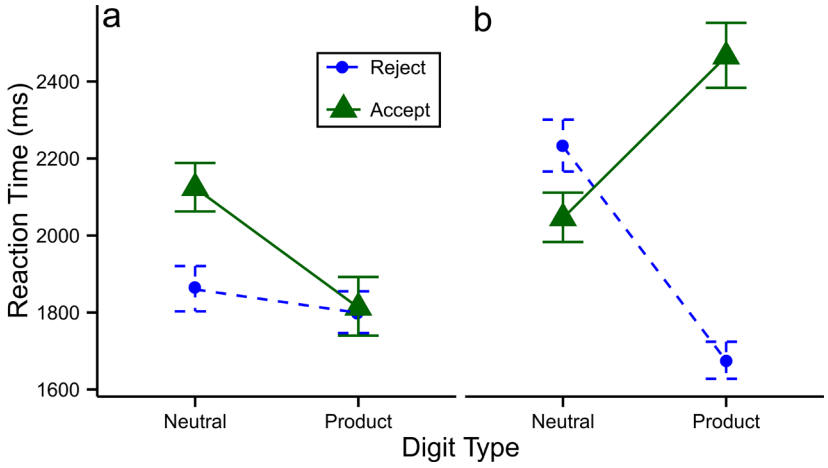

Figure 5. Mean reaction times by condition in the Sentence Verification task separated by Prompt Type and into (a) Implicative and (b) Non-Implicative trials. Error bars represent one standard error of the mean.

with participants being faster to accept correct prompts on Neutral vs. Product trials, ΔM = 1766 ms, $t(47) = 5.204$, Holm-adjusted $p < .001$, $d = .751$, and faster to reject incorrect prompts on Product vs. Neutral trials, ΔM = 1412 ms, $t(47) = 9.266$, Holm-adjusted $p < .001$, $d = 1.337$.

Linear Mixed Models

We tested for associations between achievement level and Sentence Verification task performance via linear mixed models (Table 7). Kenward-Roger approximated degrees of freedom for the Sentence Verification models were sufficiently high (lowest = 249) for *t* to practically converge to the standard normal distribution, so we instead

present *z*-tests for coefficients. Model 1 included random intercepts for each participant along with the same fixed factors as the full repeated measures ANOVA. Because of the small number of items per condition, random effects for item were not included (estimates of other parameters were similar with or without these random effects). Fixed-effect results were similar to those from the repeated measures ANOVA. In contrast to the ANOVA findings, however, the Context × Digit Type interaction now emerged as significant, β = 0.126 trials/s, $z = 5.11$, $p < .001$.

Random slopes for all main effects were then added to the model, significantly improving the fit according to a likelihood ratio test, $\chi^2(9) = 30.54$, $p < .001$, Δ conditional $R^2_{GLMM} = .028$. However, the random slope for Context was highly correlated with the random slope for Digit Type, $r = –0.93$, indicating that the model may be over-specified. The random slope for Context exhibited the least variability and was not significantly different from zero, $\chi^2(4) = 7.36$, $p = .118$, Δ conditional $R^2_{GLMM} = .003$. Model 2 therefore excluded this term (see Table 7). This suggests that the main effects of Prompt Type and Digit Type vary by *participant*, but the effect of Context does not.

In contrast to the Number Matching task, associations between math scores and contextual sensitivity *were* observed for the Sentence Verification task. Fluency was more strongly correlated with trial speed ($r = .16$) than was ACT Math ($r = .09$), so it was entered at the Model 3 stage (the reverse order led to similar conclusions). Math Fluency rate (correct answers/minute) was positively related to speed in the Sentence Verification task, with each additional problem correct per minute associated with a 0.004 trials/s increase in speed, $z = 3.30$, $p = .001$. Additionally, a significant Fluency × Context interaction emerged, with each additional correct response on math fluency yielding an increased speed differential of 0.0017 trials/s in non-implicative trials versus implicative trials, $z = 2.01$, $p = .044$. Higher-order interactions of Fluency with the Sentence Verification conditions were not significant, *p*s > .250, indicating a lack of evidence that Fluency moderates individual differences in the types of contextual sensitivity displayed in the interactions between condition variables.

The addition of ACT Math and Reading to the model provided further explanatory power. The final model for the Sentence Verification task, Model 4 (see Table 7), revealed several effects of ACT Math and Reading. Adding these terms significantly improved fit over Model 3, Kenward-Roger $F(4, 84.6) = 5.26$, $p < .001$, Δ marginal $R^2_{GLMM} = .028$. Both ACT Reading and Math interacted with Digit Type, with effects that were roughly of equal magnitude but opposite direction. No higher-order three-way interactions of the achievement measures with the condition variables were found, *p*s > .250. The predictive contribution of Math Fluency was relatively independent from the ACT measures, with little change in regression weights between Models 3 and 4.

Figure 6(a) shows the interaction of ACT Math and Digit Type from Model 4 by displaying the predicted RT across the range of ACT Math scores present in our sample (16–36), with all other variables held constant at their means. For participants with lower ACT Math scores, RTs on Neutral and Product trials differed only slightly, but the highest scoring participants showed an advantage of about 300 ms for responses to Product trials versus Neutral trials. Conversely, Figure 6(b) shows that participants with *lower* ACT Reading scores showed a substantial advantage for Product trials over Neutral trials, but there was little difference for higher-scoring participants.

After adding all significant condition and individual difference terms to the model, we examined whether there remained any evidence of unexplained individual variability in the interactions between variables. There was no evidence of participant-specific differences in the slopes of two-way interactions for Model 4, $\chi^2(22) = 20.68$, $p > .250$, Δ conditional $R^2_{GLMM} = .016$, and only marginal evidence of potential individual differences when Model 4

Table 7

Linear Mixed Effects Regression Weights of Inverse RT (trials/s) on the Sentence Verification task (n = 73)

	Model							
	(1)		(2)		(3)		(4)	
Effect	**β**	***SE***	**β**	***SE***	**β**	***SE***	**β**	***SE***
Digit Type	0.029*	0.017	0.029*	0.017	0.029*	0.017	0.050	0.076
Prompt Type	–0.056***	0.019	–0.055***	0.020	–0.055***	0.020	–0.056***	0.020
Context	–0.088***	0.017	–0.088***	0.017	–0.011	0.042	–0.012	0.042
Digit Type × Prompt Type	0.059**	0.026	0.059**	0.025	0.058**	0.025	0.059**	0.025
Digit Type × Context	0.130***	0.025	0.130***	0.024	0.120***	0.024	0.130***	0.024
Prompt Type × Context	0.110***	0.026	0.100***	0.025	0.100***	0.025	0.100***	0.025
Digit Type × Prompt Type × Context	–0.300***	0.036	–0.300***	0.036	–0.300***	0.036	–0.300***	0.036
Fluency rate					0.004***	0.001	0.004***	0.001
Context × Fluency rate					–0.002**	0.001	–0.002**	0.001
ACT Math[a]							–0.004	0.004
ACT Reading[a]							0.013***	0.003
Digit Type × ACT Math[a]							0.006**	0.002
Digit Type × ACT Reading[a]							–0.006***	0.002
Constant	0.530***	0.019	0.530***	0.019	0.340***	0.066	0.120	0.110
Akaike Inf. Crit.	–474		–487		–493		–506	
Bayesian Inf. Crit.	–418		–403		–397		–387	
Marginal R^2_{GLMM}	.059		.060		.081		.110	
Conditional R^2_{GLMM}	.303		.327		.324		.328	

Note. All variables are uncentered. Model 1 includes random intercepts nested within participants. Models 2, 3, and 4 include random intercepts and random slopes for Digit Type and Prompt Type, nested within participants. Estimated random effect parameters for Models 2, 3, and 4 are similar. Reference levels were *Reject* for Prompt Type, *Non-implicative* for Context, and *Neutral* for Digit Type. All *p* values are based on *t*-tests using the Kenward-Roger approximation for degrees of freedom. Marginal R^2_{GLMM} estimates variance accounted for by fixed effects; conditional R^2_{GLMM} estimates variance accounted for by fixed and random effects.

[a]Includes 6 participants with missing ACT Scores imputed from SAT scores.

*$p < .1$. **$p < .05$. ***$p < .01$.

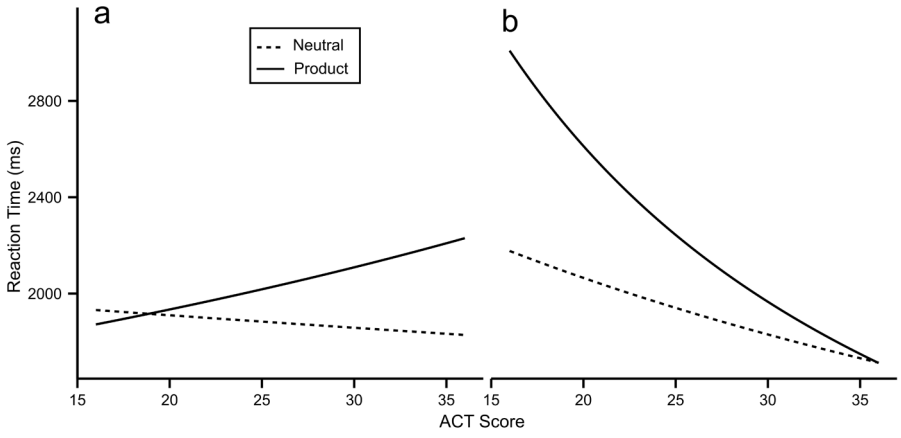

Figure 6. Modeled Interaction between Context for Products and (a) ACT Math and (b) ACT Reading on the Sentence Verification task.

was compared with a model that included random slopes for all possible condition interactions, $\chi^2(30) = 41.01$, $p = .087$, Δ conditional $R^2_{GLMM} = .029$.

Discussion

Cognitive science has demonstrated that automatized cognitive processes, including arithmetic, can be modulated by context (e.g., Spellman, Holyoak, & Morrison, 2001). This includes the effects of semantic misalignment on arithmetic in priming paradigms or word problems (Bassok et al., 2008; Fisher & Bassok, 2009), at least when arithmetic demands are fairly explicit (e.g., when a plus sign appears between digits). Our findings suggest that semantic misalignment is more complex than previously noted. Obligatory addition is affected by factors beyond *categorical* misalignment, and the *direction* of semantic modulation may change depending on the degree or type of misalignment. Whereas Bassok and colleagues found diminished priming with misaligned noun sets, facilitation effects from some misaligned noun pairs may have cancelled out interference effects of noun pairs that were only modestly misaligned with addition. We found analogous interference and facilitation effects for full sentences, depending on whether multiplication was implicated, and an unexpected *facilitative* effect when contexts were very semantically misaligned. These complex behavioral findings raise new questions about the role of semantic misalignment in contextualized numerical cognition and the integration of contextual and numerical processes more broadly. Automatized arithmetic may confer decision-making advantages even in apparently non-arithmetic contexts.

When Does Automatic Arithmetic Interfere With Correct Rejection?

Our Number Matching task showed that both the LeFevre interference (Figure 1a) and Bassok semantic alignment effects (Figure 1b) persist in the complete absence of computational notation (e.g., +). Our participants were slower to correctly reject non-matching digits on sum versus neutral trials for the categorically aligned condition (the LeFevre effect), but not for categorically misaligned conditions (when collapsed), replicating earlier work (the Bassok effect, cf. Figure 3a, Figure 1b). Our effect size for the Digit Type × Context for Sums interaction ($\eta^2 = .005$) was similar to Bassok et al.'s Experiment 1($\eta^2 = .008$).

We hypothesized that semantic misalignment lies on a continuum, with more misaligned noun pairs suppressing obligatory arithmetic to a greater degree than less misaligned nouns (Figure 7, top). When the misalignment conditions were examined separately (Figure 7, bottom), sum trials from the Number Matching task provided partial support for this hypothesis. When misaligned noun sets combined concrete and intangible nouns (the MCI condition), however, participants were *faster* to reject non-matches on sum trials even when compared to their performance on sum trials of the Misaligned Concrete-Concrete (MCC) condition. Our two misaligned conditions were clearly not equivalent; they differed in how they modulated obligatory arithmetic. Only when they were combined could we replicate Bassok et al.'s finding (2008).

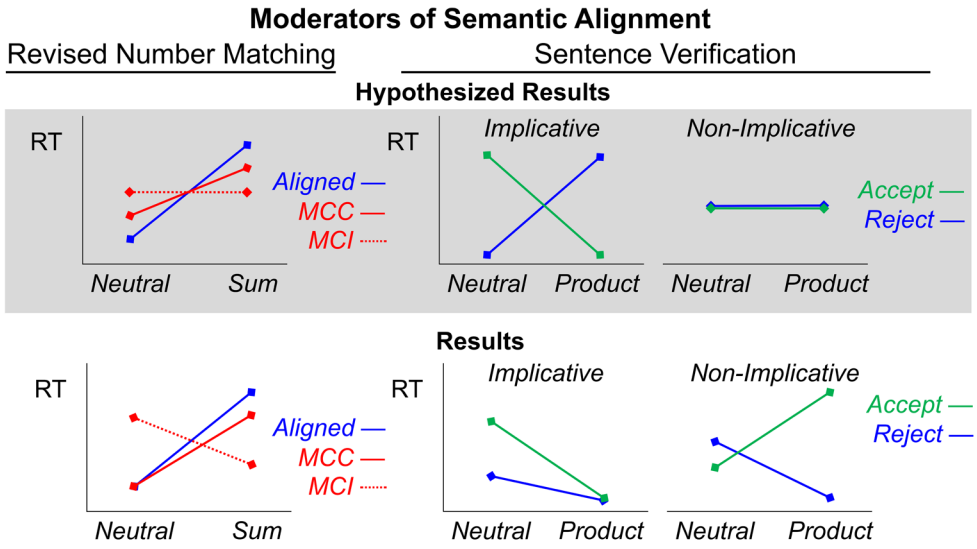

Figure 7. Priming effects for rejection trials on our revised Number Matching task and on acceptance and rejection trials on our Sentence Verification task. These effects contrast with our originally hypothesized results reported in Figure 1c, repeated here for comparison.

When Does Obligatory Arithmetic Facilitate Correct Rejection?

Our experiment provides evidence that a wholly different effect may occur in specific misaligned conditions, counter to the Bassok effect. In the Misaligned Concrete-Concrete condition, *slower* rejection of non-matching digits during *sum* versus *neutral* trials suggests that *some* arithmetic interference occurred during sum trials. However, the reverse occurred for MCI trials, with *faster* rejecting of non-matching digits during *sum* versus *neutral* trials. These differences across misaligned conditions may have emerged because we controlled for potential confounds that may underlie the lack of interference in Bassok et al.'s (2008) misaligned trials: We included only commonly enumerated nouns in our study and delineated two distinct misalignment conditions. Even if our participants simply viewed intangible nouns (e.g., myths, tactics) as less readily enumerable than concrete nouns, this would not explain the observed facilitation. This explanation is testable using an intangible noun only condition (Misaligned Intangible-Intangible), which we did not include in the present study.

Another possibility is that participants engage in a rapid, efficient, strategic rejection in the MCI condition. Results suggest that participants automatically added in all conditions, but perhaps obligatory arithmetic *assists* performance on select trials. We deliberately designed the MCI condition to be maximally unsupportive of addition, assuming that combining concrete and intangible nouns is less plausible or logical than combining even misaligned concrete nouns. (For example, combining goats and phones may be more plausible than combining goats and tactics.) But we did not anticipate that this extreme semantic misalignment may trigger an *expectation* that the sum must be an *incorrect* response to such an extent that misalignment facilitates immediate recognition (and thus rejection) of the (improbable) sum, rather than suppressing obligatory arithmetic. In contrast, *neutral* targets (which are not sums, and thus are not obvious incorrect matches) require direct comparisons and thus longer RTs. This is what we found.

Additional evidence for strategic use of semantic misalignment comes from recent ERP studies on a different type of sentence verification task (Guthormsen et al., 2016). Participants in that study saw sentences describing addition

(e.g., “Twelve bats plus two caves equals fourteen”) and responded whether the statement was “acceptable” (“Yes”/ “No” but deliberately undefined). ERP responses to the onset of the *second* noun (e.g. “caves”) in the sentence were examined for the presence of a P600 effect associated with encountering semantic anomalies. The authors found that some participants *rejected* semantically misaligned but mathematically correct statements (such as the prior example), and that others *accepted* such statements. Notably, the former group of participants responded to the second noun with a P600 effect, whereas the latter group did not. Although we did not observe individual differences on our Number Matching task (but did for the Sentence Verification task), if our participants strongly recognized a semantic anomaly between concrete and intangible nouns, this may facilitate correctly rejecting non-matching sums, whereas the semantic anomaly between categorically misaligned concrete nouns was insufficient for this judgment. Seeing a sum in such anomalous circumstances may be “uncanny” enough that participants can quickly make the “wise” choice to reject it as a non-match. Further ERP research could clarify this explanation by investigating the presence and strength of P600 effects for MCC and MCI nouns in the absence of explicit addition.

Analogous facilitation effects may also explain our Sentence Verification Task findings. In this task, participants read a sentence that either did or did not implicate multiplication, and then judged if a prompt that followed the sentence was likely to be true or false. When multiplication was *not* implicated in the initial sentence (i.e., on non-implicative trials), participants were *slower* to accept true prompts if the prompt contained products of the digits appearing in the sentence (compared to neutral digits), and they were much faster to correctly *reject* incorrect statements containing products versus neutral digits (Figures 5 and 7). Like the sums in the MCI Number Matching condition, viewing a product *inhibited* correctly accepting and *facilitated* correctly rejecting a prompt sentence when the semantic context of the target sentences did not support arithmetic. This pattern was not repeated in the implicative condition; RTs did not differ when correctly rejecting false statements regardless of whether statements contained a product.

Our findings additionally point to the importance of context beyond the semantic alignment of the nouns that accompany numbers. On the Sentence Verification trials where multiplication was not implicated, the same interaction was observed even on trials where the same unit appeared both in the cue and the target sentences. This suggests that participants react to the broader context of the cue sentence and not only to the semantic alignment of the nouns associated with the cue numbers.

Individual Differences

Are these alignment effects subject to individual differences? We found subtle evidence in the Sentence Verification task only, and equally subtle associations with arithmetic fluency scores. Fluency is ostensibly a measure of speed when correctly answering problems in a highly implicative context (an explicit arithmetic task), so it is intriguing that participants with higher math fluency scores made especially efficient use of information in non-implicative contexts. This suggests that fluency may be partially a matter of choosing operations accurately. Products within prompts interacted with math and reading achievement in opposite directions: Participants with *higher* Math or *lower* Reading ACT scores responded more quickly to neutral prompts versus product prompts; there were no differences in RTs between neutral and product prompts for participants with low Math or high Reading ACT scores. Arithmetic products may be more salient among persons with higher math achievement, which may interfere with integrating mathematically ambiguous contextual cues. Conversely, lower reading achievement may slow integration of these components simply due to more labored comprehension. This cost of math achievement echoes findings that high-numeracy participants sometimes are more negatively influenced by numerical framing

when making decisions (Peters et al., 2006). Our study adds to this finding by showing that reading achievement may index aspects of comprehending sentences with numbers, and that arithmetic fluency may provide special advantages in contexts that do not elicit arithmetic processing.

The theoretical bases for individual differences on the Sentence Verification task vary from well-documented individual differences underlying numerosity judgments (e.g., Halberda, Ly, Wilmer, Naiman, & Germine, 2012), and individual differences in cognitive control and its effect on conflict adaptation (Hsu & Novick, 2016). It is not clear why we did not replicate individual differences in LeFevre interference (LeFevre & Kulak, 1994) on our Number Matching task, especially given that individual differences emerge in ERP responses to semantic alignment (Guthormsen et al., 2016). Perhaps subtle task differences linked to the presence or absence of an arithmetic operator as the fixation point affect detection of individual differences; Price, Mazzocco, and Ansari (2013) showed individual differences in the automaticity with which young adults respond to arithmetic *computations*, not just triplet sets of numbers. It is also possible that detecting individual differences may require more difficult tasks and/or more sensitive measures (such as ERP responses). Our findings on the Sentence Verification task illustrate the need to better understand the individual differences in semantic misalignment and contextual sensitivity that have been theorized to be important for word problem solving and the development of mathematical cognition (Martin & Bassok, 2005; Mazzocco, Chan, & Sera, 2016). Finally, a more diverse sample may elicit individual differences in our Number Matching task, clarifying the connections between the Number Matching and Sentence Verification tasks and their differing levels of contextual richness.

Conclusion

How do these findings apply to everyday situations wherein numbers appear in diverse arithmetic and non-arithmetic contexts? Bassok et al. (2008) showed that people do not compute when it is illogical to do so. We show this, too, but we also show that sometimes people *do* compute when it is illogical, and that this outcome may have costs or benefits depending on whether an arithmetic operation is implicated and appropriate. The lack of any context implicating arithmetic may itself be information that some individuals seem to exploit.

Notes

i) We do not focus on mechanistic explanations for the semantic alignment effects, and thus do not attempt to resolve this issue. Following from Bassok (e.g., Bassok et al., 2008), we use the term “suppression” to refer to the phenomenon of the diminished LeFevre effect and not as an accepted mechanistic explanation.

Funding

This work was supported by a Grant-in-Aid of Research, Artistry and Scholarship from the University of Minnesota Office of the Vice President for Research to MM.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors would like to thank Chun Hei Li and Taylor Praus for dedicated assistance with data collection, scoring, and entry; Dr. Sashank Varma for early discussions concerning experimental designs; Dr. Panayiota Kendeou for input on developing our text stimuli; Drs. Sashank Varma and Keisha Varma for invaluable support by sharing their respective lab spaces for

recruitment and testing; Ella Coben for assistance in developing the noun triplets; members of the Early Math and Numeracy Lab who provided feedback on the development and piloting of our stimuli and protocols; and the Center for Cognitive Science for providing resources necessary for recruitment and data collection. The authors also thank the editor and two reviewers for feedback on an earlier version of the manuscript. MM conceived the study. MM and EB designed the study and stimuli. EB programmed all data collection protocols and oversaw the study logistics and data collection and, with input from MM and LR, analyzed the data. EB wrote the paper, with MM and LR. NS contributed to data collection, scoring and entry, and preparation of the manuscript.

References

ACT, Inc. (2014). *The ACT technical manual*. Iowa City, IA, USA: Author.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. doi:10.1016/j.jml.2007.12.005

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12-28. http://revistas.usb.edu.co/index.php/IJPR/article/view/807

Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology, 35*(2), 99-134. doi:10.1006/cogp.1998.0675

Bassok, M., Pedigo, S. F., & Oskarsson, A. T. (2008). Priming addition facts with semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(2), 343-352. doi:10.1037/0278-7393.34.2.343

Besner, D., Stolz, J. A., & Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychonomic Bulletin & Review, 4*(2), 221-225. doi:10.3758/BF03209396

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA, USA: Sage.

Campbell, J. I., & Graham, D. J. (1985). Mental multiplication skill: Structure, process, and acquisition. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 39*(2), 338-366. doi:10.1037/h0080065

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4), 447-464. doi:10.1093/llc/fqq018

de Vaan, L., Schreuder, R., & Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *Mental Lexicon, 2*(1), 1-23. doi:10.1075/ml.2.1.02vaa

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University, 73*(2), 24-32.

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

Guthormsen, A. M., Fisher, K. J., Bassok, M., Osterhout, L., DeWolf, M., & Holyoak, K. J. (2016). Conceptual integration of arithmetic operations with real-world knowledge: Evidence from event-related potentials. *Cognitive Science, 40*(3), 723-757. doi:10.1111/cogs.12238

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences of the United States of America, 109*(28), 11116-11120. doi:10.1073/pnas.1200196109

Hsu, N. S., & Novick, J. M. (2016). Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. *Psychological Science, 27*(4), 572-582. doi:10.1177/0956797615625223

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution, 5*(9), 944-946. doi:10.1111/2041-210X.12225

LeFevre, J.-A., Bisanz, J., & Mrkonjic, L. (1988). Cognitive arithmetic: Evidence for obligatory activation of arithmetic facts. *Memory & Cognition, 16*(1), 45-53. doi:10.3758/BF03197744

LeFevre, J.-A., & Kulak, A. G. (1994). Individual differences in the obligatory activation of addition facts. *Memory & Cognition, 22*(2), 188-200. doi:10.3758/BF03208890

Mazzocco, M. M. M., Chan, J. Y.-C., & Sera, M. (2016). Contextual sensitivity and the large number word bias: When is bigger really more? In A. Henik (Ed.), *Continuous issues in numerical cognition: How many or how much* (pp. 81-103). London, United Kingdom: Academic Press.

Martin, S. A., & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word-problem solving and equation construction tasks. *Memory & Cognition, 33*(3), 471-478. doi:10.3758/BF03193064

Mather, N., & Woodcock, R. W. (2001). *Examiner's manual: Woodcock-Johnson III tests of achievement*. Itasca, IL, USA: Riverside Publishing.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133-142. doi:10.1111/j.2041-210x.2012.00261.x

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447. doi:10.1037/1082-989X.8.4.434

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*(5), 407-413. doi:10.1111/j.1467-9280.2006.01720.x

Price, G. R., Mazzocco, M. M. M., & Ansari, D. (2013). Why mental arithmetic counts: Brain activation during single digit arithmetic predicts high school math scores. *The Journal of Neuroscience, 33*(1), 156-163. doi:10.1523/JNEUROSCI.2936-12.2013

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510-532. doi:10.1037/0033-2909.114.3.510

Rivera, S. M., Reiss, A. L., Eckert, M. A., & Menon, V. (2005). Developmental changes in mental arithmetic: Evidence for increased functional specialization in the left inferior parietal cortex. *Cerebral Cortex, 15*(11), 1779-1790. doi:10.1093/cercor/bhi055

Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory & Cognition, 29*(3), 383-393. doi:10.3758/BF03196389