Verbal interference suppresses exact numerical representation

Citation: Frank, Michael C., Evelina Fedorenko, Peter Lai, Rebecca Saxe, and Edward Gibson. “Verbal Interference Suppresses Exact Numerical Representation.” Cognitive Psychology 64, no. 1–2 (February 2012): 74–92.

As Published: http://dx.doi.org/10.1016/j.cogpsych.2011.10.004

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/102467

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Verbal interference suppresses exact numerical representation

Michael C. Frank[a], Evelina Fedorenko[b], Peter Lai[b], Rebecca Saxe[b], Edward Gibson[b]

[a] *Department of Psychology, Stanford University*
[b] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

Abstract

Language for number is an important case study of the relationship between language and cognition because the mechanisms of non-verbal numerical cognition are well-understood. When the Pirahã (an Amazonian hunter-gatherer tribe who have no exact number words) are tested in non-verbal numerical tasks, they are able to perform one-to-one matching tasks but make errors in more difficult tasks. Their pattern of errors suggests that they are using analog magnitude estimation, an evolutionarily- and developmentally-conserved mechanism for estimating quantities. Here we show that English-speaking participants rely on the same mechanisms when verbal number representations are unavailable due to verbal interference. Followup experiments demonstrate that the effects of verbal interference are primarily manifest during encoding of quantity information, and—using a new procedure for matching difficulty of interference tasks for individual participants—that the effects are restricted to verbal interference. These results are consistent with the hypothesis that number words are used online to encode, store, and manipulate numerical information. This linguistic strategy complements, rather than altering or replacing, non-verbal representations.

Keywords:

1. Introduction

How does knowing a language affect the way you perceive, act, and reason? Do differences between languages cause systematic differences in the cognition of their speakers? Questions about the relationship between language and thought are among the most controversial in cognitive science

(Boroditsky, 2001; Davidoff et al., 1999; Gentner & Goldin-Meadow, 2003; Gumperz & Levinson, 1996; Levinson et al., 2002; Li & Gleitman, 2002; Pinker, 1994; Rosch Heider, 1972). Theoretical proposals concerning the nature of this relationship run the gamut from suggestions that individual languages strongly influence their speakers cognition (Davidoff et al., 1999; Levinson, 2003; Whorf, 1956) to suggestions that there is no causal relationship between speakers' language and their cognition (Fodor, 1975; Li & Gleitman, 2002; Pinker, 1994), with a number of more moderate proposals falling between these extremes (Gentner, 2003; Kay & Kempton, 1984; Slobin, 1996).

Recently, across the domains of color, number, navigation, theory of mind, and object individuation, there has been a convergence of empirical results addressing this question. In each of these domains, meaningful cognitive differences have been demonstrated between people who have words for particular concepts and those who don't, either because their language does not encode those concepts (Frank et al., 2008; Gordon, 2004; Pica et al., 2004; Pyers & Senghas, 2009; Winawer et al., 2007) or because they haven't yet learned the relevant words (de Villiers & de Villiers, 2000; Le Corre et al., 2006). These group differences are mitigated or disappear entirely when the people who do know the relevant words cannot access these words (for example, when they are required to occupy their verbal resources with interfering material or when tasks are speeded) (Hermer-Vazquez et al., 1999; Ratliff & Newcombe, 2008; Newton & de Villiers, 2007; Winawer et al., 2007).

These similarities across domains point towards a unified account of the relationship between language and cognition that falls midway between the two theoretical extremes of strong interaction and no interaction. On the one hand, the data suggest that languages do change the cognition of their speakers: they help their speakers accomplish difficult cognitive tasks by creating abstractions for the efficient processing and storage of information. On the other hand, the data also suggest the hypothesis that these abstractions complement rather than replace pre-existing non-verbal representations. When linguistic abstractions are temporarily inaccessible, language users seem to fall back on the representations used by other animals, children, and speakers of languages without those abstractions.

This relationship—the use of language online for encoding—has been referred to in a number of ways in the previous literature. Kay & Kempton (1984) follow Whorf (1956) in describing cognition as having two tiers: "one, a kind of rock-bottom, inescapable seeing-things-as-they are (or at least as human beings cannot help but see them) and a second, in which the metaphors implicit in the grammatical and lexical structures of language cause us to classify things in ways that could be otherwise (and are otherwise for speakers of different languages)." Gentner & Goldin-Meadow (2003) and Frank et al. (2008), emphasizing the way that linguistic representations augment cognition, refer to this view as "language as a toolkit" or "cognitive technology." Dessalegn & Landau (2008) refer to this as the "momentary" hypothesis, emphasizing that the role of language is online rather than permanent. All of these accounts posit that tasks like verbal interference temporarily disable this second tier, leading speakers to perform tasks in ways that are shared cross-culturally. The goal of the current experiments is to test this prediction for numerical cognition: that experienced number users under verbal interference should perform numerical tasks in the same way as people with no number words.

1.1. Numerical cognition and language

As a case study of the relationship between language and thought, number has a key advantage over other domains: the pre-linguistic mechanisms for representing numerical information are relatively well-understood (Cantlon et al., 2009; Carey, 2009; Dehaene, 1997). Numerical cognition in infants and non-human animals is characterized by two distinct systems (Feigenson et al., 2004). The parallel-individuation ("object file") system is used to track the identity of a few (up to three or four) discrete objects. In contrast, the analog magnitude system is used to represent large, approximate quantities. Analog magnitude estimation operates over arbitrarily large quantities, but the error in the estimate increases in proportion to the size of the set being estimated (a constant coefficient of variation, or COV; see Appendix) (Shepard, 1975; Whalen et al., 1999).

In the absence of words for numbers, even human adults appear to rely on these core numerical systems. For example, adults' estimates of quantity show the same systematic errors as those of infants and pigeons when the sets are presented too rapidly to be counted (Whalen et al., 1999). When a culture has no words for number, the same profile is observed even for slower presentation rates, as documented in two Amazonian groups, the Pirahã (Gordon, 2004) and the Mundurukú (Pica et al., 2004). Evidence from the Pirahã have been used to support a further claim, as well. Gordon reported that the Pirahã language had only three numerical words, roughly corresponding to the concepts of "one," "two," and "many," and that Pirahã people were unable to perform simple numerical matching tasks, including creating a set that was the same size as a target set using one-to-one correspondence. He interpreted these results as evidence for a strong Whorfian claim: without language for number, he argued, the Pirahã had no concept of exact quantity.

Our own recent results slightly alter this picture. We reported that Pirahã actually has no words for exact quantities; the words previously glossed as exact numerals ("one" and "two") apparently are comparative or relative terms (Frank et al., 2008). In addition, we showed that the Pirahã were able to succeed in simple one-to-one matching tasks, suggesting that they did understand the principle of exact, one-to-one correspondence, even for large sets. However, our results were similar to those reported by Gordon in that the Pirahã made systematic errors on matching tasks that required memory for exact quantities. Thus, current evidence from the Pirahã suggests that the ability to remember and manipulate exact quantities, but not the concept of exact correspondence, relies on language.

Another set of recent results seems to conflict with this account, however. Butterworth et al. (2008) investigated a group of numerical tasks with children ages 4–7 who had grown up speaking Warlpiri or Anindilyakwa, two native Australian languages. Both languages have some number morphology (e.g. singular, dual, plural in Warlpiri) and Anindilyakwa has a base-5 number system, though it is not in heavy use. To test for effects of language on numerical cognition, Butterworth and colleagues compared data from tasks on cross-modal matching, addition, and number memory (as well as a sharing task in which performance depended on learned strategies) to control data from 4–5 year-old English-speaking children from an urban environment. They found that all three groups performed comparably in all tasks, with age emerging as the major factor driving performance. A second study suggested that spatial grouping strategies might provide a non-linguistic alternative to exact enumeration (Butterworth & Reeve, 2008). Across both studies, they interpreted this lack of a language effect as suggesting that language was not the key factor in the development of enumeration abilities, contra work with the Pirahã and Mundurukú.

One salient issue in comparing these data to the previous Amazonian results was the pattern of errors shown by the English-speaking participants. In two out of three tasks in Butterworth et al. (2008), the young English-speaking children—like the Warlpiri and Anindilyakwa children—made responses consistent with approximate number use. In fact, across both sets of studies performance by the English-speakers in all tasks was strikingly low: for example, performance in simple addition problems like 3+1 was below 50%, and it dropped to about 20% in problems like 5+3. This pattern suggests that even if the children had mastered the count list, they were not using their exact number knowledge to succeed and may even have had trouble understanding the tasks. The Australian data thus do not provide a strong test of what role language plays in establishing exact number concepts because even numerate participants in their study failed to use exact number concepts.

Nevertheless, the Butterworth et al. account makes a strong prediction that contradicts the theoretical accounts of the online role of language in cognition that are described above. If language is not crucial in establishing exact number concepts either developmentally or in the moment, adult speakers of English should be able to perform exact numerical tasks under verbal interference. In contrast, if language is necessary for online storage of exact numbers, English speakers should rely on analog magnitude estimation when they cannot use linguistic resources. Our experiments evaluate this prediction.

Several previous studies have investigated numerical tasks under conditions designed to suppress or circumvent the use of language, but they used paradigms that were at least partially verbal in nature. Logie & Baddeley (1987) conducted a detailed investigation of the effects of interference tasks on verbal counting. They found that verbal suppression via rapid repetition of the word "the" caused participants to make errors in counting, while simply tapping a finger or listening to speech caused far fewer. They concluded that the articulatory-phonological loop (Baddeley, 1987) was strongly implicated in counting. In a followup study, Trick (2005) investigated the effects of simple and complex verbal suppression and motor interference tasks (repeating one letter/tapping one finger or alternating between two letters/two fingers) and found substantial effects of both complex verbal and complex motor tasks on enumeration. Because neither of these studies provided evidence about the variability of participants' errors, it is not possible to determine whether participants were making use of analog magnitude estimation (as would have been shown by a constant COV).

Two studies have investigated the relationship between counting and COV. Whalen et al. (1999) showed participants Arabic numerals and then asked them to press a key so quickly that they could not verbally enumerate the number of times they did so. Cordes et al. (2001) used a similar task but asked participants either to perform a verbal suppression task (repeating "the") or to count as fast as they could. These studies found that participants showed a constant COV under speeded response and verbal suppression, signaling a reliance on analog magnitude estimation. In contrast, when counting they showed a decreasing COV, suggesting a pattern of binomial errors (caused by errors in the correspondence between the verbal count list and their key presses; see the Appendix for more details on how different COV trends can be derived from different numerical mechanisms or strategies).

These studies provide direct evidence that verbal suppression affects counting performance. However, they investigated explicitly verbal tasks (either counting a quantity or translating a numeral into a set of actions). It is still unknown whether participants from a numerate culture who are under verbal interference will rely on analog magnitude estimation in completely non-verbal tasks. To evaluate this question, we conducted three experiments. Experiment 1 tests English speakers' performance under verbal interference when tested in completely non-verbal tasks identical to those used with the Pirahã. We find that, while English speakers primarily use ad-hoc non-linguistic strategies in simpler matching tasks, in the most demanding tasks they rely exclusively on analog magnitude estimation. Experiment 2 investigates how verbal resources facilitate the storage and manipulation of quantity information, testing whether verbal interference impairs verbal encoding of quantity information, or whether it has an equal effect on retrieving quantity information once it is encoded. We find that encoding, as opposed to retrieval, is differentially affected by verbal interference. Experiment 3 tests whether it is specifically the inaccessibility of linguistic resources that forces a reliance on the approximate number system by comparing the effects of matched verbal and spatial interference tasks. We find that language interference produces both a greater degree and different pattern of impairment in numerical performance.

2. Experiment 1

In our first experiment, we compared the previously reported matching task performance of Pirahã participants—who lack words for exact numbers entirely—to new data from English-speakers under verbal interference—for whom number words were temporarily unavailable. We performed the same set of tasks that we used with the Pirahã with English-speaking participants in Boston, MA while these participants verbally shadowed radio news broadcasts (repeating words out loud as they were heard over headphones) to block access to number words (Hermer-Vazquez et al., 1999; Newton & de Villiers, 2007). Although this design would normally require a control group that performed the tasks without shadowing, pilot testing convinced us that, for numerate adults, performance would be at ceiling without the presence of a concurrent verbal task like shadowing.

Previous work with the Pirahã revealed significant variability in performance across different matching tasks (the set of tasks used here and in Frank et al., 2008 is shown in Figure 1). All these tasks require the participant to construct a line of objects with the same quantity as those shown by the experimenter. Simple one-to-one matching tasks between aligned sets were easiest for all participants; tasks where participants had to match a rotated or hidden set were harder; and the most difficult task was the "nuts-in-a-can" task, where the experimenter's set was placed in an opaque cup one at a time. This pattern of results is easily interpreted: of these tasks, the nuts-in-a-can task is the only one where the full set of objects was not visible, denying participants the ability to use the physical extent of the experimenter's set as an extra cue for matching their own set. We thus believe that nuts-in-a-can is the task that best measures the effects of counting on participants' performance, because it most effectively prevents the use of non-linguistic strategies other than magnitude and duration estimation. We thus predict that the match between Pirahã participants' performance and that of the English-speakers under verbal interference should be closest on this task and on the one-to-one match tasks. In the nuts-in-a-can task we predict a flat COV, as observed with the Pirahã; in the one-to-one match task, we predict ceiling effects.

2.1. Methods

2.1.1. Participants

We recruited 35 participants from MIT and the surrounding community; our participants varied in age from 18 to 50, approximately matching the range of ages in our Pirahã population (though exact matching was of course impossible because the Pirahã could not report their ages). We excluded data for a task on the basis of individual error rates greater than three standard deviations above task mean. This criterion resulted in the exclusion of one participant in four of the five tasks (with no individual participant excluded in more than one task).

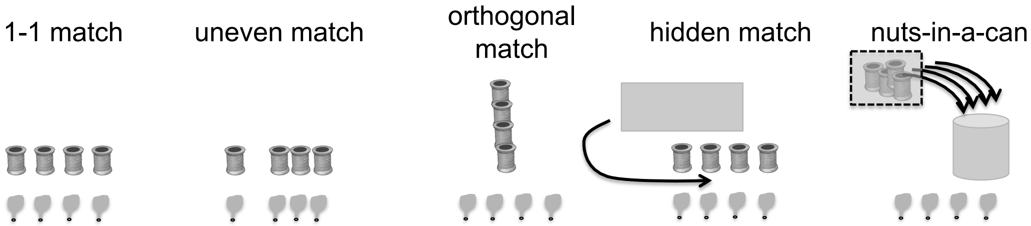

Figure 1: Schematic of each of the tasks used in Experiment 1. Participant begins verbal shadowing, experimenter places spools of thread from a larger set, and participant attempts to place the same quantity of balloons from their own set.

2.1.2. Procedure

Participants were first familiarized with the verbal shadowing task: they were instructed to listen to short clips from the Radio News Corpus (Ostendorf et al., 1995) and to repeat the words spoken by the announcer as quickly as possible. After their performance was judged to be fluent by the experimenter and they reported that they were comfortable with the task, they were given instructions for the matching tasks.

Each participant completed five matching tasks (Figure 1), in the following order: a one-to-one matching task, an uneven matching task, an orthogonal matching task, a hidden matching task, and a nuts-in-a-can task. The order of tasks was kept constant across participants; tasks were ordered from easiest to most difficult, as in work with the Pirahã, in order to allow participants to get used to the shadowing task and the response format. Each matching task required the participant to observe some quantity of spools of thread and to put out a line of uninflated balloons exactly matching the quantity of spools that they saw (these items were chosen because they were the same stimulus items we used with the Pirahã). For each task, participants were tested once on each quantity from 4–12 (in one of two random orders), and the number of balloons they put out was recorded by the experimenter. This procedure resulted in 9 trials per task for a total of 45 trials per participant. No feedback was given.

In the one-to-one and uneven matching trials, the experimenter placed the spools one by one in a line running from the participant's left to their right. The spools were evenly spaced in the one-to-one task and broken randomly into smaller groups of one to four (still in the same order) in the uneven task. The line of balloons placed by participants was parallel to the line placed by the experimenter, so the participants could simply place balloons in one-to-one correspondence with spools to succeed in the task. The orthogonal matching task was identical to the one-to-one task except that the line of spools ran from closer to the participant to further away, rather than from left to right. The hidden match task was identical to the one-to-one task except that the line was hidden by the experimenter after the spools were placed (by placing a manila folder in front of the spools). Finally, in the nuts-in-a-can task, the experimenter placed spools one by one into an opaque cup.

On each trial, the experimenter would begin by starting the audio (which the participant listened to over headphones). Once the participant had begun shadowing, the experimenter placed the spools one by one in the task configuration. Once the experimenter had finished, the participant began placing balloons to indicate quantity. When finished placing balloons, the participant indicated that the trial was finished by pressing a key to end the audio.

2.2. Results

Figure 2 shows the distribution of responses across groups and tasks (Pirahã data are re-plotted from Frank et al. 2008, while Figure 3 shows the proportion of correct responses and coefficient of variation. For both populations, the one-to-one and uneven matching tasks were easiest, with performance close to ceiling. Likewise, for both populations the nuts-in-a-can task was hardest; although the English-speakers were more accurate than the Pirahã, both groups made significant and systematic errors. However, the performance of the two groups diverged on the hidden and orthogonal match tasks. While these tasks were only slightly more difficult than the one-to-one and uneven match tasks for the English-speakers, they were far more difficult than the one-to-one and uneven match tasks for the Pirahã.

To test for differences between tasks and groups, we fit a single logistic mixed model to the entire dataset (Gelman & Hill, 2006). This model attempted to predict participants' performance on individual trials on the basis of fixed effects of task (one-to-one, uneven, orthogonal, hidden match, or nuts-in-a-can), group (Pirahã vs. English), quantity being estimated, and a random intercept term for each participant. In this and all other models, quantity was modeled as a continuous numerical predictor. The use of the mixed model allowed us to use the fact that all participants completed all

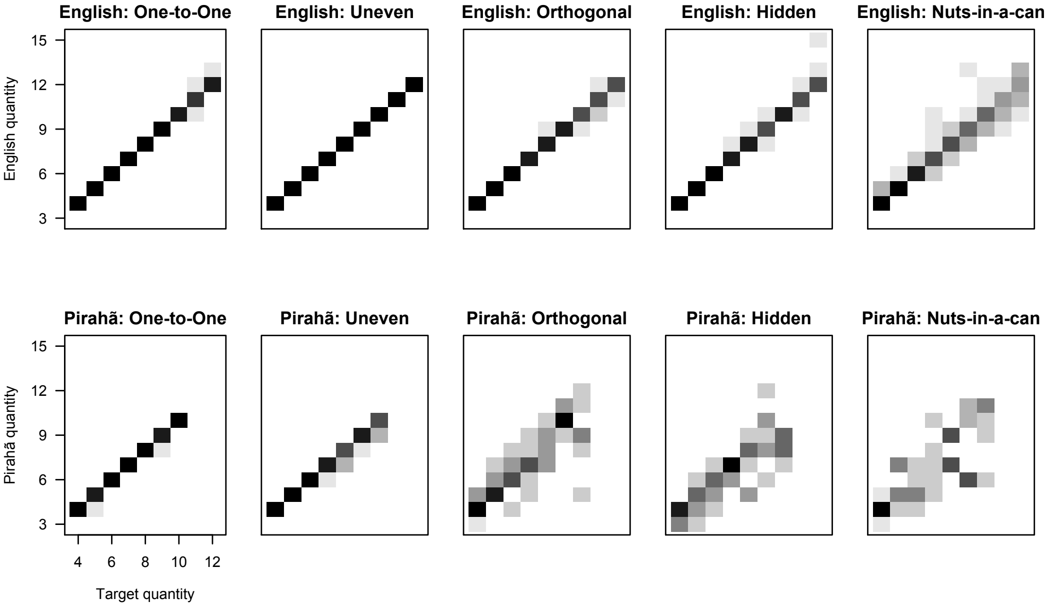

Figure 2: Responses of English (top rows) and Pirahã (bottom rows) groups in each of the five tasks tested in Experiment 1 (English) and Frank et al. (2008) (Pirahã). Plots show the probability distribution over response quantities for each quantity that was tested (darker = higher probability).

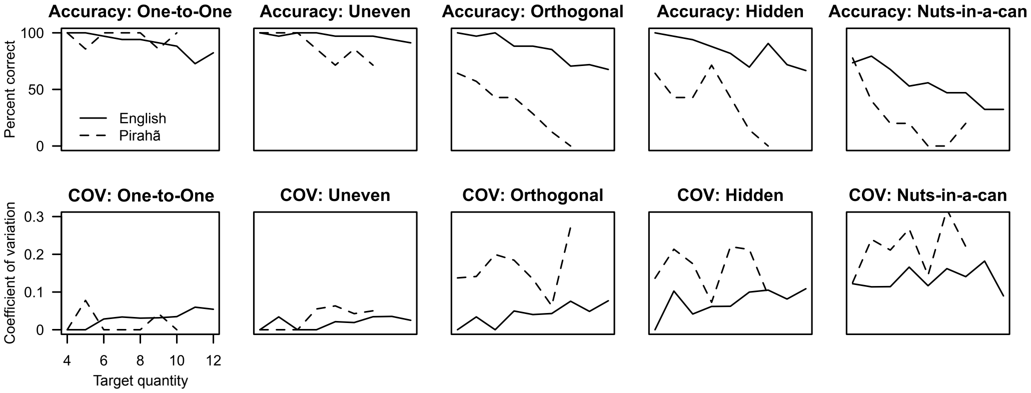

Figure 3: Accuracy and coefficient of variation (mean/standard deviation) by task for each group. Horizontal axis shows the quantity of objects presented by the experimenter. English speakers' results are plotted with a solid line, Pirahã results are shown by a dashed line.

Table 1: Coefficient estimates and 95% confidence intervals produced by posterior simulation for a logistic mixed model predicting participants' performance. Each coefficient represents an estimate of the effects of that task on participants' accuracy, expressed in logistic units. A single coefficient was fit for the effects of target quantity (size of target set) on accuracy.

		Coefficient estimate	Lower 95% CI	Upper 95% CI
English	One-to-one	5.71	4.96	6.54
	Uneven	6.94	5.97	7.93
	Orthogonal	5.07	4.38	5.83
	Hidden	4.97	4.24	5.69
	Nuts-in-a-can	3.14	2.55	3.75
Pirahã	One-to-one	8.35	4.94	23.80
	Uneven	4.94	3.93	6.26
	Orthogonal	1.88	1.09	2.70
	Hidden	2.05	1.30	2.82
	Nuts-in-a-can	1.56	0.62	2.43
Target quantity		-0.37	-0.43	-0.30

Table 2: Summary statistics comparing Pirahã and English-speaking participants' results.

Percent correct refers to the total proportion of trials of a particular type that were correct. Mean error is the average difference between the mean of participants' responses and the target value (small values indicate that participants' estimates were correct across the group, despite being variable for any individual trial). Mean COV refers to the mean coefficient of variation for a task (see text); a smaller COV indicates more precise estimates. COV slope is the slope of the best linear fit to the coefficient of variation, as a function of quantity. COV r^2 and p-value are the r^2 and p values for this linear fit.

Group	Task	% correct	Mean error	Mean COV	COV slope	COV r^2	COV p-value
English	One-to-one	0.91	0.04	0.03	0.007	0.83	< 0.01
	Uneven	0.97	0.02	0.02	0.003	0.33	0.11
	Orthogonal	0.86	0.05	0.04	0.008	0.67	< 0.01
	Hidden	0.84	0.20	0.08	0.009	0.47	0.04
	Nuts-in-a-can	0.54	0.18	0.14	0.001	0.01	0.81
Pirahã	One-to-one	0.96	0.04	0.02	-0.002	0.03	0.71
	Uneven	0.89	0.12	0.03	0.010	0.64	0.03
	Orthogonal	0.39	0.50	0.16	0.002	0.01	0.86
	Hidden	0.43	0.58	0.15	-0.003	0.01	0.83
	Nuts-in-a-can	0.32	0.39	0.21	0.009	0.08	0.53

tasks to factor out participant-level variability from our effects of interest, as well as allowing us to compare both between- and within-group coefficient estimates.

Although we initially included a term for the interaction of task and quantity in the model, we pruned this term from the model as it added a large number of degrees of freedom without significantly increasing fit (likelihood comparison between models resulted in a test statistic of $\chi^2(9) = 9.59$, $p = .38$). We next used posterior simulation from the model to compute confidence intervals for each coefficient. Coefficients indicate relative amounts of change in the probability of a correct response in the task; their magnitudes are not directly interpretable, but represent changes in the probability of a correct response at a particular performance level (as in logistic regression more generally). Estimates of each coefficient and 95% confidence intervals for each group and task are shown in Table 1. These coefficients reflect the size of the effect on accuracy of participating in each task. For example, for English speakers, the coefficient estimate for one-to-one matching was 5.71, while for nuts-in-a-can it was 3.14, indicating a considerable drop in overall accuracy.

English speakers made fewer errors across nearly all tasks. The posterior distribution of the difference between coefficient estimates for Pirahã and English participants differed for the uneven ($p < .05$), orthogonal ($p < .001$), hidden ($p < .001$), and nuts-in-a-can ($p < .001$) tasks, but not for the one-to-one matching task ($p > .1$). However, the magnitude of the difference between groups varied from task to task. The posterior difference between groups was significantly smaller for the nuts-in-a-can task than it was for either the orthogonal ($p < .01$) or hidden match ($p < .05$) tasks.

We additionally compared the coefficient of variation for each participant across groups. We take the dependence of COV on quantity as a measure of the approximate number system; while COV averaged across quantities provides a general measure of overall error.

COV is normally computed for a particular quantity by dividing the standard deviation of estimates of a particular quantity by the mean of the estimates (Gordon, 2004). Although participants did not contribute multiple data points for each quantity, we approximated COV c for individual participants' n different responses $r_1 ... r_n$ on target quantities $t_1 ... t_n$ by $c = 1/n \sum_i \sqrt{(t_i - r_i)^2}/t_i$. This approximation is derived as follows: since $c = \sigma/\mu$, for the j measurements of estimates of a particular quantity i,

$$c_i = \frac{\sqrt{n^{-1}\sum_j (r_j - \bar{r_i})^2}}{n^{-1}\sum_j r_j}. \tag{1}$$

In the case where there is only a single measurement, this simplifies to $c_i = \frac{\sqrt{(r_j - \bar{r_i})^2}}{r_j}$. We approximate $\bar{r_i}$—the average judgment of a quantity—as t_i (the quantity being estimated); this approximation is justified by the extremely high correlation (r values are often $> .99$) between these two values in the aggregate data across experiments and participants. We then average c_i across quantities to produce a mean COV.

Using this approximation, COV did not significantly differ in the one-to-one match task ($t(47) = -.77$, $p = .45$) or the uneven match task ($t(47) = -1.43$, $p = .17$). In contrast, in the three other tasks, orthogonal match ($t(47) = -4.01$, $p = .001$), hidden match ($t(47) = -3.76$, $p = .002$), and nuts-in-a-can ($t(41) = -2.53$, $p = .03$), the Pirahã had a higher COV. Table 2 shows within-group summary statistics. We used linear regression to predict COV as a function of quantity in order to determine whether there was a significant increasing or decreasing linear trend in the COV of each group for each task. Both the Pirahã and the English speakers showed a flat COV in the nuts-in-a-can task, according to our predictions. The Pirahã showed a significant trend in COV for only the uneven match task. In contrast, the English speakers showed a significantly increasing COV for the one-to-one match task, the hidden match task, and the orthogonal match task, as well some hints of a trend in the uneven match task.

2.3. Discussion

Congruent with the "language as a tool" and "momentary" views, we predicted that, when verbal resources were unavailable, English speakers would fall back on analog magnitude estimation in the same tasks as the Pirahã. Because the nuts-in-a-can task provides a particularly good test of whether speakers can use an exact strategy, we focus primarily on this task. Both groups showed a constant COV in this task, indicating exclusive reliance on the analog magnitude estimation system, contra predictions from the Butterworth et al. (2008) account. In addition, both groups also showed near-ceiling levels of performance in the one-to-one and uneven matching tasks, demonstrating the ability to perform exact correspondences without number words. Despite their radically different cultural and linguistic backgrounds, both groups appeared to rely on the same non-linguistic systems in the absence of language for number.

Nevertheless, two differences between the two groups' performance require some elaboration. First, the English speakers showed higher mean accuracy and lower mean COV than the Pirahã. This group difference could be due to a variety of factors. The English speakers included undergraduates at a selective American university and thus were sampled from the overall population differently than the Pirahã participants, a theoretically uninteresting difference. However, it is also possible that the English speakers' differential cultural experience with mathematics and other uses of exact numerosity led to their relatively more precise representation of analog magnitude. Although our current data do not directly speak to this possibility, it is an intriguing avenue for further research (Halberda et al., 2008).

The second difference concerned the ease of the hidden and orthogonal match tasks for the English speakers relative to the Pirahã. Why did these two tasks cluster with the easier one-to-one and uneven match tasks for the English speakers but with the more difficult nuts-in-a-can task for the Pirahã? We use the English speakers' COV to distinguish three possible explanations for this difference (see Appendix for details). First, if participants were forced to rely exclusively on analog magnitude estimation in these tasks, then performance should have shown a similar pattern to the nuts-in-a-can task: steadily increasing errors in proportion to the target set size, resulting in a constant COV across quantities. Second, if participants had exclusively used a verbal strategy—counting via systematic avoidance or circumvention of verbal shadowing—their COV would show a decrease. Finally, if participants had used an ad-hoc, failure-prone strategy to supplement analog magnitude estimation, their COVs would show an increase.

The COV of English speakers' judgments increased with the quantity being estimated, consistent with the use of ad-hoc strategies. During debriefing, participants reported trying to form mental subsets of the objects in the hidden match task (”subitizing”), trying to co-register objects one by one across sets in the orthogonal match task, and occasionally trying to count the objects despite interference. We therefore propose that the English speakers performed better than the Pirahã (and better than expected via analog magnitude estimation) on these two tasks by using a mixture of explicit exact strategies and analog magnitude estimation.

3. Experiment 2

Our next experiment focused in on the nuts-in-a-can task. As discussed above, we believe this task provides the cleanest test of performance, since participants cannot rely on other visual strategies to succeed. Because the finding of a constant COV for English speakers in this task was the most important result of Experiment 1, we were interested in replicating this finding. In addition, we were interested in whether verbal interference differentially affected encoding (creating a mental representation of a set's quantity) or retrieval (constructing a set using a stored quantity). Roberson & Davidoff (2000) suggested that verbal interference during encoding might impair categorization in a color task, but the data were equivocal, perhaps due to the short duration of the experiment's encoding phase. The longer encoding phase—the time during which the experimenter placed the spools of thread into the container—in the nuts-in-a-can task thus provides an opportunity for a more direct comparison of interference during encoding to interference during retrieval.

3.1. Methods

3.1.1. Participants

We recruited 15 participants from MIT and the surrounding community, who participated for compensation as part of a larger group of studies.

3.1.2. Procedure

As in Experiment 1, participants were first familiarized with the verbal shadowing task. They were then each tested in the nuts-in-a-can task in three separate conditions, counterbalanced for order. The *Encoding & Retrieval* condition was identical to the nuts-in-a-can task in Experiment 1. In the *Encoding* condition, participants performed the verbal shadowing task while the experimenter placed the spools in the cup, then the experimenter ended the verbal shadowing task and the participants made their responses with no secondary task. In the *Retrieval* condition, participants watched the experimenter place the spools in the cup with no verbal shadowing, then made their responses while performing the verbal shadowing task.

3.2. Results

The distribution of participants' responses is plotted in Figure 4 and participants' accuracy and COV are plotted in Figure 5. Summary statistics are

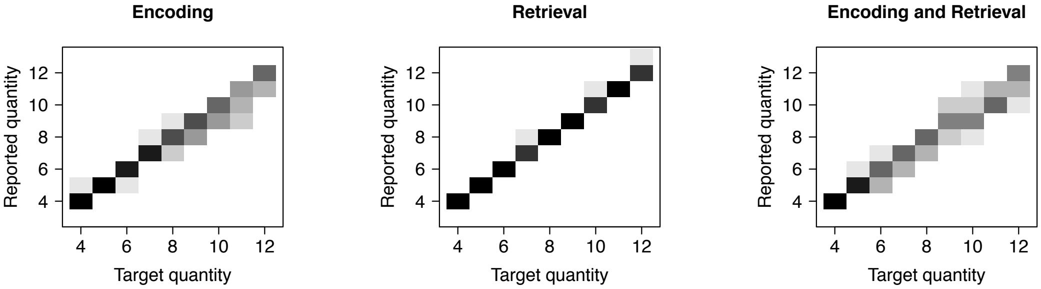

Figure 4: Participants' responses in the three conditions of Experiment 2. Plots show the probability distribution over response quantities for each quantity that was tested (darker = higher probability).

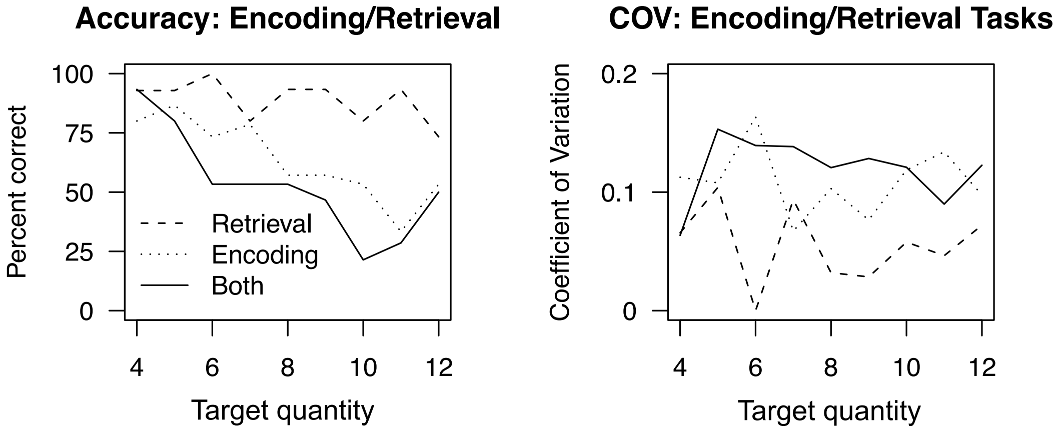

Figure 5: Accuracy and coefficient of variation (mean/standard deviation) for each condition in Experiment 2. Results from the retrieval, encoding, and encoding/retrieval (both) conditions are plotted with dashed, dotted, and solid lines respectively.

Table 3: Summary statistics comparing participants' results across conditions in Experiments 2 and 3.

Group	Condition	% correct	Mean error	Mean COV	COV slope	COV r^2	COV p-value
Expt. 2	Encoding	0.66	0.30	0.11	-0.003	0.08	0.47
	Retrieval	0.87	0.13	0.07	-0.004	0.18	0.26
	Both	0.53	0.27	0.12	0.000	0.00	0.91
Expt. 3	Spatial	0.86	0.01	0.04	-0.009	0.75	0.01
	Verbal	0.72	0.16	0.07	0.000	0.00	0.96

provided in Table 3. We again constructed a generalized linear mixed model predicting participants' error as a function of condition (encoding, retrieval, or encoding/retrieval) and target quantity. We tested whether adding interaction terms for the three conditions with target quantity significantly added to model fit but found that they did not ($\chi^2(2) = .98$, $p = .61$). The lack of condition by target quantity interactions is due to the use of a logistic model to fit the binomial accuracy data in this task. Because of the shape of the logistic curve, the slope of the accuracy function will be different within different quantity ranges even without an interaction term. An interaction would be predicted only if there were a difference in the pattern of errors in different conditions (for example, if participants were not using the analog magnitude system in one condition).

We again used posterior simulation to compute confidence intervals for each coefficient estimate. We found that performance in the encoding condition was only marginally better than performance in the encoding/retrieval condition ($\beta = .52$, $p < .1$), while performance in the retrieval condition was considerably better ($\beta = 2.22$, $p < .0001$). Corroborating these findings, we saw that average participant COVs (calculated as in Experiment 1) in the encoding and encoding/retrieval conditions were different from those in the retrieval condition (paired $t(14) = 3.34$, $p < .01$ and $t(14) = 3.73$, $p < .01$) but did not differ from one another ($t(14) = .91$, $p = .38$). Mean COV in each of the conditions showed no linear trend, indicating that participants' pattern of errors did not provide evidence for mechanisms other than analog magnitude estimation for any condition.

3.3. Discussion

Our results from the Encoding / Retrieval condition of Experiment 2 replicated the pattern of results in the nuts-in-a-can condition of Experiment 1. In addition, we found that correctly encoding the quantity of spools under verbal interference was far more difficult for participants than placing the correct quantity of balloons under interference once that quantity was encoded. Although both phases of the task included the same actions (placing each object in a larger set, one by one) and took approximately the same amount of time, performance was disrupted when the set's cardinality could not be verbally encoded. If participants were able to count and linguistically encode the target set's quantity, it was much easier for them to check their set's cardinality against that of the target set.

In all three conditions, however, we observed a relatively constant COV, suggesting the operation of the analog magnitude system. If linguistically-encoded exact number representations were being used in the Retrieval condition, why did we not observe a decreasing COV? We first note that although a linear term did not reach significance in this condition, there was some hint of a negative slope in the Retrieval COV. However, this effect was likely masked by some analog (estimation-based) noise during retrieval. We speculated that a more precise paradigm with a larger amount of data at each quantity might show this effect more clearly. This issue motivated the development of the computer-based paradigm used in Experiment 3.

4. Experiment 3

Our goal in our final experiment was to test whether verbal interference specifically—rather than cognitive load more generally—was responsible for the use of the approximate number system in the nuts-in-a-can task in Experiments 1 and 2.

Because the results of Experiment 2 suggested that the effect of verbal interference was primarily manifest during encoding, and because the manual nature of the tasks used in Experiments 1 and 2 limited the number of trials we could conduct (setting up each trial and manually initiating verbal interference took considerably longer than a comparable computerized display), we used a computerized version of the nuts-in-a-can paradigm in Experiment 3. We asked participants to view a computerized display in which dots sequentially appeared in the center of the screen and to "place" a matching quantity of dots by pressing the space bar to make dots appear in a row on the screen. We then used this paradigm to contrast directly the effects of verbal and non-verbal interference on number judgment.

The next experimental challenge was creating matched verbal and non-verbal interference tasks. As demonstrated by Trick (2005), the complexity of the verbal interference task has a significant effect on participants' error rates. Thus, we chose not to use repetition of a single word, the interference task that was used by previous researchers in this domain (Cordes et al., 2001; Logie & Baddeley, 1987). Yet verbal shadowing, the interference task we used in Experiment 1, has no directly analogous non-verbal correlate. It is both easier to learn—most participants require only a minimum of training or explanation—and more complex than comparable non-verbal shadowing tasks—conveying far more information than a simple clapped rhythm (Newton & de Villiers, 2007; Hermer-Vazquez et al., 1999). Therefore, in order to match our verbal and non-verbal interference tasks as closely as possible, we made use of verbal and spatial short-term memory tasks.

Previous investigations of verbal and non-verbal interference used tasks which caused a similar mean decrement in performance on a secondary control task (one outside the domain of interest). This method does not ensure that all individuals performed equivalently on both interference tasks. For example, Newton & de Villiers (2007) conducted a pilot study in which they asked participants to perform a number of interference tasks as they performed a visual orientation judgment task; they found that the average difficulty of their rhythm imitation and verbal shadowing tasks did not show significant differences. Likewise, Lupyan (2009) used a visual search task with a set of verbal and non-verbal memory tasks and found no difference in task performance between the two interference tasks. Nevertheless, even in the absence of a statistically significant result, there could have been differences between proficiency on the two tasks for individual participants. For example, in our pilot testing we found that rhythm imitation was far easier for participants with musical training than for those who had never performed this sort of task, which could lead to large differences in the effects of shadowing depending on population differences.

In order to circumvent this issue, we designed a paradigm that ensured that interference tasks were matched for each individual participant on the amount they distracted from an independent target task. The target task we chose was serial visual search (Lupyan, 2009). To ensure that the interference tasks were matched, we created an adaptive staircase paradigm that adjusted the difficulty threshold for the two tasks for each participant until the same empirical level of performance was reached. In this first phase, a trial consisted of: a memory stimulus presentation, either a sequence of consonants or a sequence of spatial locations in a grid); a search stimulus presentation, in which an L was either present or not present among an array of Ts; a 2-alternative forced choice search response; and a memory response. The difficulty of the search task was held constant across trials, while the number of items in the memory sequence was adjusted according to participants' performance.

In the second phase of the experiment, the level of difficulty in each interference task was set to equate participants' performance across tasks and held constant through this phase. These matched interference tasks were then paired with a numerical task analogous to the nuts-in-a-can task, in which participants were shown a set of dots that flashed briefly around the middle of the screen and then had to produce a linear array of the same cardinality by pressing the space bar to increase the number of dots shown in a line. The logic of this task is then as follows: because the difficulty of the interference tasks was matched for each participant, any differences in quantity manipulation that we observe in this paradigm are not due to differences in difficulty between the two interference tasks (or how well they interfere with a third, target task) but must be due to an interaction between the particular interference task and the quantity judgment task.

4.1. Methods

4.1.1. Participants

Twenty-four Stanford undergraduates participated in exchange for course credit.

4.1.2. Procedure

Stimuli were presented using custom Matlab software (version 2010a) written with the Psychtoolbox package (Brainard, 1997). The experiment was broken into two conditions, spatial interference and verbal interference. Each participant completed both conditions (order was randomized across participants). The structure of each condition was as follows. Participants first completed 60 interference/search trials, staircased to find a constant level of performance. They next completed 65 interference/quantity estimation trials.

Interference tasks consisted of the presentation of either a string of consonants appearing sequentially in the same location or a sequence of blue squares appearing sequentially in different locations of a 4x4 grid. Timing was the same for both stimuli: each item (consonant or square) was presented for 200ms and there was a 100ms interval between presentations. Presentation was rapid to prevent rehearsal between the presentation of subsequent items. Between encoding and retrieval, search trials were presented. Search trials consisted of an array of 24 letters scattered around the screen. On half of trials, all letters were Ts, on the other half, one letter was an L. Participants were instructed to press a key to indicate whether there was an L present or not. After a search response was made, participants indicated their response in the interference trial by typing the consonants in any order on the keyboard or by clicking locations in the grid with the mouse in any order. Feedback about the correctness of a response was given immediately after the end of a trial and indicated whether participants had made an error on either of the tasks, both, or neither.

The adaptive staircase began with two items; if participants gave a correct response to both search and interference tasks on two subsequent trials, the quantity of items presented in the next interference task trial was increased by one. if they made an error on either or both tasks, the quantity of items in the interference task was decreased by one. Participants were informed of the structure of the staircase prior to testing.

In quantity estimation trials, the number of interference items was fixed as the average number of items for that participant during the last 25 trials of their staircase, rounded up (pilot testing indicated that the staircase usually converged after approximately 25–35 trials). The interference items were first presented, then a string of black dots were shown sequentially in locations jittered slightly from the middle of the display. This jitter was small, so that spatial density was not a cue, but was enough to provide some spatial individuation information. Participants were prompted to create an array with the same quantity by pressing the space bar to "place" dots in a line stretching from left to right. If too many dots were placed, participants could press backspace to remove dots. Finally, participants recalled the interference items as in the staircase trials. Participants were given feedback only on the interference portion of their response, in order to encourage them to maintain their focus on the interference task. Dots were shown for 500ms each with a 250ms interval between presentations. Quantities 1–10 were each shown six times, and quantities 11–15 were each shown once in order to keep participants from noticing an explicit ceiling on the quantity distribution. Order of presentation was fully randomized for each participant. Each condition of the experiment lasted approximately 30 minutes for a total of one hour of testing.

4.2. Results

The distribution of participants' responses is shown in Figure 6, their accuracy and COV is plotted in Figure 7, and summary statistics are given in Table 3.

We first examined results from the staircase (search plus interference) trials. We found that participants did not differ on search performance or reaction time across conditions (accuracy M=.99 for spatial interference, M=.98 for verbal interference, $t(23) = .79$, $p = .44$; reaction time M=2.46 s for

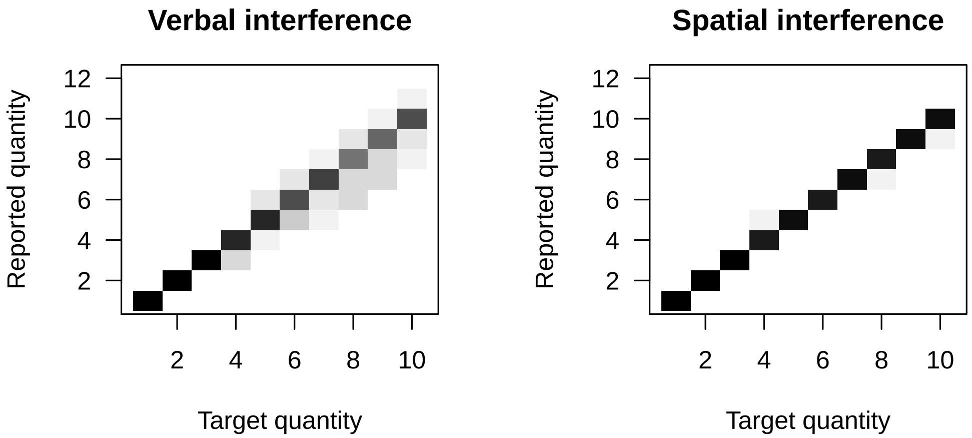

Figure 6: Participants' responses in the two interference conditions of Experiment 3. Plots show the probability distribution over response quantities for each quantity that was tested (darker = higher probability).

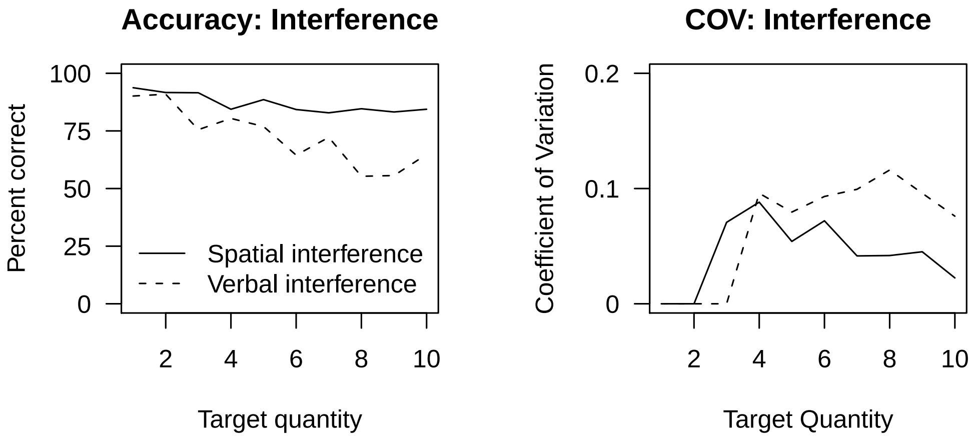

Figure 7: Accuracy and coefficient of variation (mean/standard deviation) for the two conditions in Experiment 3. Results from the spatial interference condition are plotted with a solid line, while results from the verbal interference condition are plotted with a dashed line.

spatial interference, M=2.46 s, $t(23) = -.04$, $p = .97$). The verbal task produced slightly higher interference task values after thresholding (M=4.96 for spatial interference, M=5.29 for verbal interference, $t(23) = 2.00$, $p = .06$), indicating that these tasks were not completely matched and hence it was important to use the adaptive staircase procedure (as we did) rather than fixing a single interference level for both.

The goal of our next analysis was to ascertain whether performance was comparable on quantity trials with matched degrees of verbal and spatial interference. We therefore excluded all responses in the quantity estimation task for which the response was not correct on the corresponding interference trial. This manipulation eliminated 47.53% of all trials (36.7% for spatial trials and 58.4% for the verbal trials). While this may seem like a large proportion of trials, one goal of this experiment was to ensure that interference load was exactly matched across participants and tasks. On trials where participants made errors in the interference task, we do not know the cause of these errors. In some cases they may have encoded the interference string (providing the interference effect even though the interfering material was not recalled correctly), but they may also have failed to encode this material, allowing them to focus on the numerical task for that trial. We also did not include the small set of trials at cardinalities above 10, as these were only included in order to avoid participants inferring that 10 was the ceiling of responses in the task.

We fit a generalized linear mixed model to correct trials on cardinalities 1–10 including effects of condition (spatial vs. verbal) and target quantity, as well as an interaction of the two (and a random effect of participant). We found that an interaction term significantly increased model fit ($\chi^2(1) = 6.73$, $p = .009$), so it was retained in our simulations. We again used posterior inference to find the empirical distribution of coefficient values. In the case of this model, we found no main effect of condition ($\beta = -.24$, $p = .24$), but a main effect of target quantity ($\beta = -.11$, $p < .001$) and an interaction of target quantity and condition ($\beta = -.15$, $p < .01$). [1]

We next calculated the coefficient of variation for each participant in

[1] We repeated all of these analyses with all trials included (as opposed to excluding trials in which there was an error on the interference task). The pattern of results did not change qualitatively or in level of significance: we still found no main effect of condition ($\beta = -.26$, $p = .48$) and significant effects of target quantity ($\beta = -.11$, $p < .01$) and the interaction of target quantity and condition ($\beta = -.14$, $p < .01$).

each condition. Because analog magnitude estimation is widely assumed to be masked by the operation of the small number system, we were primarily interested in COV for quantities 4–10. Although we now had 5 measurements per quantity per participant, due to the removal of incorrect interference trials, the number of observations for each quantity was small and variable. We thus used the same approximation of COV as described in our methods for Experiment 1. In addition, in order to remove outliers whose large magnitude would compromise COV estimates (typos or trials where participants simply guessed), we omitted datapoints where responses were more than two standard deviations away from the mean response (approximately 2.7% of data). Exactly equivalent numbers of outliers were omitted from each condition, and approximately twice as many were omitted for the quantities 1–5 as 5–10. This stringent outlier criterion was necessary because the COV is calculated on the basis of root mean squared error and so (like other RMS-based methods, e.g. standard linear regression) is very sensitive to values that fall far from the mean.

Participants' mean COVs across quantities between the two conditions differed significantly from one another (paired $t(23) = 2.29$, $p = .03$). In addition, we found a significantly decreasing slope across quantities in the mean COV in the spatial interference condition, while the mean COV was flat in the verbal interference condition (Table 3).

4.3. Discussion

Experiment 3 was designed to test whether the nature of the interference task (in particular, verbal vs. non-verbal) would have an effect on participants' performance. To that end we designed a procedure for matching the difficulty of verbal and spatial short-term memory tasks for each participant and then used these matched tasks as interference tasks for quantity estimation. In Experiment 2 we found no interaction of quantity and condition, presumably due to the same process (analog estimation) operating in all conditions. In contrast, in Experiment 3, we saw a significant interaction, indicating that the relationship between target quantity and accuracy was different by condition (not just that one condition was harder than the other). Participants in the spatial condition were able to count (the more accurate strategy), while participants in the more difficult verbal interference condition could not count and were forced to estimate.[2]

The COV data confirmed that there was a qualitatively different pattern of errors between conditions. COVs were higher in the verbal interference condition and were constant across quantities, as they were in the comparable nuts-in-a-can conditions in Experiments 1 and 2, signaling use of the analog magnitude system. In contrast, in the spatial interference condition participants showed a low and decreasing COV, signaling counting. The Appendix describes the prediction of a decreasing COV for counting, originally from Cordes et al. (2001).

To summarize: verbal interference has a different effect on quantity encoding than matched spatial interference. This difference is likely attributable to the fact that participants could not count when they were performing a challenging verbal interference task, but had much less difficulty doing so when they were performing a difficulty-matched spatial interference task.

5. General Discussion

The goal of our experiments was to test whether, when access to number words was impaired via verbal interference, participants would use analog magnitudes to keep track of approximate quantities. Experiment 1 showed that, like the Pirahã, who have no words for numbers, numerically-savvy English speakers will also rely on analog magnitude estimation when they are prevented from using linguistic resources—though they do this only in tasks where it is difficult to use other ad-hoc strategies. Experiment 2 provided a replication of this finding and new evidence that the primary effect of language interference is on the encoding, not the retrieval, of quantity information. Experiment 3 established that this effect was specific to verbal interference, since a difficulty-matched spatial interference task produced results consistent with a linguistic strategy.

In addition to knowing how to count, the English-speaking participants in our experiments brought with them a lifetime of formal education in a

[2] In addition to matching distractor (span) tasks on their effects on a visual search task, in a separate experiment (N=20) we matched the distractor tasks on their difficulty without the intervening visual search task. The pattern of results was qualitatively quite similar, with significantly lower accuracy in the verbal interference condition compared with the spatial interference condition, and a flat COV in the verbal—but not the spatial—condition.

culture radically different from that of the Pirahã, with all of the cognitive differences associated with this different background (Scribner & Cole, 1973). Given these differences, the similarities in performance between the two groups were considerable: in the one-to-one and uneven matching tasks, both English- and Pirahã-speakers were able to put objects in exact correspondence, and in the nuts-in-a-can task, both groups relied on analog magnitude estimation. Taken together, our data support the following interpretation: the concept of "exact match" does not depend on language in either its genesis or its use, but encoding, storing, and manipulating exact quantities larger than three or four relies crucially on verbal representations, at least in the practice of the Pirahã and English speakers in our studies. These linguistic representations—when present—are used in the moment in which the operations are carried out, consistent with the "momentary" and "language as a tool" hypotheses (Dessalegn & Landau, 2008; Gentner & Goldin-Meadow, 2003; Frank et al., 2008).

What is it about having language for number that enables better performance on numerical tasks? Performing simple tasks like the ones tested here requires storage or manipulation of an exact quantity. But even for these tasks—and for a range of others—it is necessary to have both an enumeration routine (a count list or tally system) and a representation of the product of that routine (the numeral that was counted to or the resulting tally mark). Without some type of enumeration routine routine, the cardinality of a set can only be estimated (Spaepen et al., in press; Flaherty & Senghas, under review). Likewise, without a numeral or some other physical or mental summary representation of the quantity, the quantity information cannot be retained. Thus, language enables numerical performance by providing both a representation and a routine for manipulating exact quantities.

Nevertheless, the use of non-linguistic strategies in Experiment 1 and the broader literature on other methods of enumeration and calculation suggest that language is only one source of representations and routines for keeping track of exact quantities. For example, a tally board, a knot system, or an abacus can be used to keep track of an exact quantity without language (although the physical representation must be present at the time of recall for these methods) (Menninger, 1969). The use of a mental representation of an abacus, however, provides one example of a general non-linguistic representation of exact number that is in widespread use (Frank & Barner, 2011; Hatano, 1977). Thus, language is not the only way to manipulate exact number; it is simply a flexible, powerful, and common method.

All human beings (and many other species) share some core, prelinguistic numerical capacities. Languages which have words for representing exact quantities—and other consistent representation systems—allow their users to transcend these capacities and attain genuinely new numerical abilities (Le Corre et al., 2006; Carey, 2009). These novel abilities depend on both having new representations (large exact numbers) and learning new routines for manipulating these representations (addition, division, factorization, etc.). The benefit of these abilities can be as simple as remembering quantities with higher fidelity over time or as complex as the technical innovations made possible by calculus and differential equations. But even within the domain of number, not all systems are created equal. Some count lists may be so difficult to learn or use that they impair the ability of their speakers to count large numbers of objects (Donohue, 2008; Saxe, 1999); in other cases, count lists may evolve for specific purposes, rather than the general goal of representing all exact cardinalities (Beller & Bender, 2008; Evans, 2009). The abilities enabled by exact number representations depend crucially on the structure of those representations.

As demonstrated by our current results, however, the addition of these new representations does not replace the core numerical abilities, which are still accessible and on which English-speakers still rely heavily when their verbal resources are otherwise occupied. When lexical representations of exact quantities are available, they serve as a preferred route for processing and storing numerical information. In the absence or inaccessibility of this route, the original core abilities of object individuation and magnitude estimation are still present.

Appendix A. Signatures of numerical processes in the coefficient of variation

The literature on numerical cognition provides a number of tools for characterizing performance. Of these, the most important is the *coefficient of variation* (COV). The COV c for a set of numerical responses $R = r_1...r_m$ with a target quantity n is defined as:

$$c(t) = \frac{\sigma(R)}{\mu(R)} \tag{A.1}$$

In responses to changes in n, the COV can stay constant, decrease, or increase. Previous literature has linked the relationship between COV and n to particular psychological processes (Whalen et al., 1999; Cordes et al., 2001; Gordon, 2004).

The strongest of these predictions, and one that is confirmed by a wide variety of empirical literature, is the link between a constant COV and the use of the analog magnitude system. This result follows from the Weber-Fechner law (Fechner, 1960), which states that the smallest perceptible difference dp in a stimulus s is related to the magnitude of that stimulus:

$$dp = c\frac{ds}{s} \tag{A.2}$$

When two stimuli whose perception follows the Weber-Fechner law are compared to one another, their discriminability varies as the ratio of their means. For example, in Whalen et al. (1999), individual participants were asked to produce a number of key presses equivalent to an Arabic numeral. The variability in the number of presses they produced varied as a function of the quantity they were attempting to estimate. When variability was normalized by quantity (to produce a COV), it was found to be constant, as in previous work with animals (e.g. (Platt & Johnson, 1971)).

Cordes et al. (2001) contrasted this prediction with one derived from the idea of errors made in verbal counting. They assumed that participants doing speeded verbal counting would make errors consistent with a binomial probability of error α, related to a binomial probability of independently omitting any number in the count list. This pattern of responding produces a COV that is predicted to decrease proportional to the square root of n, and the empirical data from a speeded counting task confirmed this prediction.

Why then did participants in some of our tasks show significantly *increasing* COV values as quantities increased? In investigations like those of Cordes et al. (2001) and Logie & Baddeley (1987), participants were explicitly asked to count and so they may have stayed with this strategy when it was suboptimal. In contrast, in our task, we did not specify what strategy should be used. Thus, we suggest that participants monitored their performance and switched strategies if it became clear that one had failed. For example, if participants knew that they had miscounted, misaligned, or simply lost track, a good response would be to switch to a (much faster and less effortful) estimation strategy.

We describe a mixture model that assumes that participants rely on an exact strategy until they detect that it has failed and then switch to an estimation strategy. This model is based on participants' verbal reports that they used a variety of exact strategies—including but not limited to verbal counting—to succeed in the orthogonal and hidden match tasks. For example, several participants reported that they carried out the orthogonal match task by grouping the spools into several smaller sets and then making their responses by recreating these subgroups and mentally aligning them to their positions in the larger group. Other participants said that they attempted to count the objects in the conditions where all objects were present (orthogonal and hidden match) despite the verbal interference.

We start with the supposition that if an exact strategy succeeds it has error $\epsilon = 0$. This will happen with probability proportional to $(1-\alpha)^n$, where α is the probability of success and n is the quantity being estimated. If the strategy fails, error will be normally distributed around c following Weber's law with constant k: $\epsilon \sim N(0, (ck)^2)$. The mixture of these two strategies gives

$$\epsilon \sim (1 - (1 - \alpha)^n) \cdot N(0, (ck)^2). \tag{A.3}$$

This model produces an overall increasing COV. At small quantities, errors in the exact strategy will be relatively unlikely, leading to a small probability of relying on the analog magnitude system. At large quantities, the COV will approach k. Note also that as α approaches 1, this model returns to pure analog magnitude estimation. Thus, even if participants had attempted to count in the nuts-in-a-can task, the longer duration of the task and lack of visual cues may have ensured that they lost count in the vast majority of trials due to verbal interference.

References

Baddeley, A. (1987). *Working memory*. Oxford, UK: Oxford University Press.

Beller, S., & Bender, A. (2008). The limits of counting: Numerical cognition between evolution and culture. *Science*, *319*, 213–215.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive Psychology*, *43*, 1–22.

Brainard, D. H. (1997). The psychophysic toolbox. *Spatial Vision*, *10*, 433–436.

Butterworth, B., & Reeve, R. (2008). Verbal counting and spatial strategies in numerical tasks: Evidence from indigenous australia. *Philosophical Psychology*, *21*, 443–457.

Butterworth, B., Reeve, R., Reynolds, F., & Lloyd, D. (2008). Numerical thought with and without words: Evidence from indigenous australian children. *Proceedings of the National Academy of Sciences*, *105*, 13179.

Cantlon, J., Platt, M., & Brannon, E. (2009). Beyond the number domain. *Trends in Cognitive Sciences*, *13*, 83–91.

Carey, S. (2009). *The origin of concepts*. Oxford University Press, USA.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin and Review*, *8*, 698–707.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, *398*, 203–4.

de Villiers, J., & de Villiers, P. (2000). Linguistic determinism and the understanding of false beliefs. In P. Mitchell, & K. Riggs (Eds.), *Children's reasoning and the mind* (pp. 191–228). Hove, UK: Psychology Press.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. New York: Oxford University Press.

Dessalegn, B., & Landau, B. (2008). More than meets the eye: The role of language in binding visual properties. *Psychological Science*, *19*, 189–195.

Donohue, M. (2008). Complexities with restricted numeral systems. *Linguistic Typology*, *12*, 423–429.

Evans, N. (2009). Two pus one makes thirteen: Senary numerals in the morehead-maro region. *Linguistic Typology*, *13*, 321–335.

Fechner, G. (1960). *Elemente der Psychophysik*. Leipzig: Breitkipf & Härtel.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314.

Flaherty, M., & Senghas, A. (under review). Numerosity and number signs in deaf Nicaraguan adults, .

Fodor, J. A. (1975). *The language of thought: a philosophical study of cognitive psychology*. Language and thought series. New York: Crowell.

Frank, M. C., & Barner, D. (2011). Representing exact number visually using mental abacus. *Journal of Experimental Psychology: General*, .

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, *108*, 819–824.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Gentner, D. (2003). Why we're so smart. *Language in mind: Advances in the study of language and thought*, (pp. 195–235).

Gentner, D., & Goldin-Meadow, S. (2003). Whither whorf. In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and cognition* (pp. 3–14). Cambridge, MA: MIT Press.

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, *306*, 496–499.

Gumperz, J. J., & Levinson, S. C. (1996). *Rethinking Linguistic Relativity*. Cambridge, UK: Cambridge University Press.

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement.

Hatano, G. (1977). Performance of expert abacus operators. *Cognition*, *5*, 47–55.

Hermer-Vazquez, L., Moffet, A., & Munkholm, P. (2001). Language, space, and the development of cognitive flexibility in humans: the case of two spatial memory tasks. *Cognition*, *79*, 263–299.

Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, *39*, 3–36.

Kay, P., & Kempton, W. (1984). What is the sapir-whorf hypothesis? *American Anthropologist*, *86*, 65–79.

Le Corre, M., Van de Walle, G., Brannon, E. M., & Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive Psychology*, *52*, 130–169.

Levinson, S. C. (2003). *Space in language and cognition*. Cambridge, UK: Cambridge University Press.

Levinson, S. C., Kita, S., Haun, D. B. M., & Rasch, B. H. (2002). Returning the tables: language affects spatial reasoning. *Cognition*, *84*, 155–188.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, *83*, 265–294.

Logie, R., & Baddeley, A. (1987). Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 310–326.

Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review*, *16*, 711.

Menninger, K. (1969). *Number words and number symbols: A cultural history of numbers*. Cambridge, MA: MIT Press.

Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking: Adults fail nonverbal false-belief reasoning. *Psychological Science*, *18*, 574–579.

Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). *The Boston University Radio News Corpus*. Technical Report Boston University.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an amazonian indigene group. *Science*, *306*, 499–503.

Pinker, S. (1994). *The Language Instinct*. Penguin Books.

Platt, J. R., & Johnson, D. M. (1971). Localization of position within a homogenous behavior chain: Effects of error contingencies. *Learning and Motivation*, *2*, 386–414.

Pyers, J., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science*, *20*, 805–812.

Ratliff, K., & Newcombe, N. (2008). Is language necessary for human spatial reorientation? reconsidering evidence from dual task paradigms. *Cognitive Psychology*, *56*, 142–163.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, *28*, 977–986.

Rosch Heider, E. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*, 10–20.

Saxe, G. B. (1999). Cognition, development, and cultural practices. In E. Turiel (Ed.), *Culture and Development* New Directions in Child Psychology (pp. 19–36). San Francisco, CA: Josey-Bass Publishers.

Scribner, S., & Cole, M. (1973). Cognitive consequences of formal and informal education. *Science*, *182*, 553–559.

Shepard, R. N. (1975). The internal representation of numbers. *Cognitive Psychology*, *7*, 82–138.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz, & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–95). Cambridge: Cambridge University Press.

Spaepen, L., Coppola, M., Spelke, E. S., Carey, S., & Goldin-Meadow, S. (in press). Number without a language model. *Proceedings of the National Academy of Sciences*, .

Trick, L. M. (2005). The role of working memory in spatial enumeration: Patterns of selective interference in subitizing and counting. *Psychonomic Bulletin & Review*, *12*, 675–681.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.

Whorf, B. L. (1956). *Language, thought, and reality*. Cambridge, MA: MIT Press.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*, 7780–7785.