Exact number concepts are limited to the verbal count range

Benjamin Pitt*[1]
Edward Gibson[2]
Steven T. Piantadosi[1]

January 29, 2021

1 Department of Psychology, University of California, Berkeley, CA 94720
2 Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139

Abstract

Previous studies suggest that mentally representing exact numbers larger than four depends on a verbal count routine (e.g. "one, two, three..."). However, these findings are controversial, as they rely on comparisons across radically different languages and cultures. We tested the role of language in number concepts within a single population – the Tsimane' of Bolivia – where knowledge of number words varies across individual adults. We used a novel data analysis model to quantify the point at which participants switched from exact to approximate number representations during a simple numerical matching task. The results show that these behavioral switchpoints were bounded by participants' verbal count ranges; their representations of exact cardinalities were limited to the number words they could recite. Beyond that range, they resorted to numerical approximation. These findings resolve competing accounts of previous findings and provide unambiguous evidence that large exact number concepts are enabled by language.

Introduction

Language gives humans extraordinary cognitive abilities, but its role in numerical cognition remains unresolved. Studies of human infants and non-human animals have shown that at least some numerical abilities do not depend on language. Babies, monkeys, and even invertebrates can make precise distinctions between small quantities without counting (up to about four; 1, 2) and can rapidly distinguish the numerosities of larger sets, although only roughly (3–7). Whereas the ability to represent *small exact* and *large approximate* numbers is conserved across species, the ability to represent larger numbers exactly (e.g. exactly seven) appears to be unique to humans (3; c.f. 8) and is often attributed to language (7, 9–11). Specifically, predominant accounts posit that the structure of the verbal count list (e.g. "one, two, three..."), which children learn to recite long before they understand the meanings of the number words (12–14), allows them to discover the logic of number by induction (9, 15–18; cf. 19–24).

This account draws support from studies of isolated groups with few or no words for exact quantities (11, 25–27). Specifically, two indigenous groups in the Brazilian Amazon – the Pirahã and the Mundurukú – have no words denoting large exact quantities (and in the case of the Pirahã, no words for any exact quantity, not even one; 25, 26). To test large exact number concepts in such groups without using number words, researchers have used simple numerical tasks that require only behavioral responses, often on sets of physical objects (e.g. 7 pebbles; Figure 1; 11, 25–27). Pirahã and Mundurukú adults perform well on these tasks only up to about four; for larger cardinalities, they are unable to reproduce the number of objects in a set exactly, relying instead on approximation (11, 25, 26). A similar pattern has been found in Nicaraguan Homesigners, a group of congenitally deaf adults whose language lacks a count routine (27). Across groups, the pattern is the same: People without words for large exact numbers seem unable to represent cardinalities larger than four, leading some scholars to conclude that the verbal count list "enables exact enumeration" (11).

However, these findings are difficult to interpret (20, 23, 24, 28, 29), in part because they rely on comparing across languages and cultures. Groups without exact number words (like the Pirahã) are compared, if only implicitly, to groups with productive counting systems (like Americans). Of course, isolated groups differ radically from Western, Educated, Industrialized, Rich, and Democratic (WEIRD; 30) groups in many ways besides in their knowledge of number words (e.g. 31), and any of these differences could account for the observed difference in numerical cognition (27, 32). For example, some scholars suggest that the Pirahã failed to make exact numerical matches of large sets not because they lacked the requisite linguistic resources, but because they were simply "indifferent to exact numerical equality" (20; also see 27, 29), perhaps because "keeping track of large exact quantities is not critical for getting along in Pirahã society" (28). Indeed, whereas quantification is prized in WEIRD cultures, some unindustrialized groups like the Pirahã do not track chronological age, use currency, or have units of measurement (31, 33, 34). In short, cross-cultural comparisons cannot in principle distinguish whether large exact number concepts depend on a verbal count routine or on other aspects of language and culture.

Even if these studies clearly established a causal role for language in large exact number concepts, it remains unclear what role that would be. Some accounts posit that the verbal count list is instrumental both for *inducing* the principles of number (e.g. Hume's principle: one-to-one correspondence guarantees numerical equivalence; 7, 35) and for *using* those principles to construct representations of specific cardinalities (e.g. exactly seven; 15). Alternatively, the verbal count list may be necessary for inducing the logic of number only, which people could then use to enumerate large sets whether or not the corresponding verbal symbols were available to them. Previous cross-cultural studies cannot distinguish between these possibilities because they test numerical abilities only at the extremes. In principle, the failure of the Pirahã (and other groups without large exact number words) to represent large exact numbers could be due to a lack of the requisite number principles, number words, or both.

To date, few studies have tested the role of number words in large exact number concepts without comparing across language groups (36), and the results are difficult to interpret. In a group of MIT undergraduates, verbal interference impaired performance on some numerical tasks more than a spatial control task, suggesting a functional role for language in representing large exact numbers. However, despite verbal interference, participants performed well on two other tests of large exact number representations, including the orthogonal matching task, complicating interpretation of the results. (Even if verbal interference had caused unambiguous impairments in participants' numerical abilities, it is unclear whether such an effect would generalize beyond this highly-specialized sample of WEIRD adults, given their decades of dependence on verbal number symbols.) In another study, US children overwhelmingly failed to make exact numerical matches of large sets, but this failure is difficult to interpret given their imprecision in a task that only required one-to-one matching of objects (35). In sum, previous studies do not clearly establish whether or how language influences the representation of large exact numbers.

Here we addressed these inferential challenges by testing the relationship between number words and number concepts in the Tsimane', a group of unindustrialized farmer-foragers indigenous to the Bolivian Amazon (37), who differ importantly from previously studied populations. Unlike the Pirahã, Mundurukú, and Nicaraguan Homesigners, the Tsimane' have a fully productive system of number words in their language. Yet, unlike adults in WEIRD cultures, Tsimane' adults exhibit considerable variation in their knowledge of the verbal count list; many Tsimane' adults can count indefinitely, but some do not know words above 10, others falter at 12, etc. This variability allowed us to compare verbal and numerical abilities across *individuals*, rather than across groups. It also allowed us to test the relationship between verbal and numerical abilities not just at the extremes, but at many intermediate levels. To determine which large numbers participants could represent exactly, and which numbers they could only approximate, we used a novel statistical analysis to model participants' behavioral responses in an orthogonal matching task. This model uses the known psychophysical properties of numerical estimation to determine the set size at which participants switched from exact to approximate number representations. By comparing this *switchpoint* to participants' highest verbal counts, we tested whether people need a system of number symbols (like those in the verbal count list) in order to represent large exact numbers. If they do (16, 17), then we should find not only that these abilities are correlated, but that one systematically exceeds the other; participants' highest verbal counts should place an *upper bound* on their numerical representations, allowing them to make exact matches only within the limits of their verbal count range. Alternatively, if number words are necessary for discovering the logic of number but not for deploying it (or not at all, e.g. 21, 23), then participants' numerical representations should sometimes exceed their verbal count ranges. Unlike in previous studies, here the relationship between verbal counting and numerical reproduction cannot be attributed to broad cultural or linguistic differences, since our participants shared the same culture, language, and in many cases lived in the same small community.

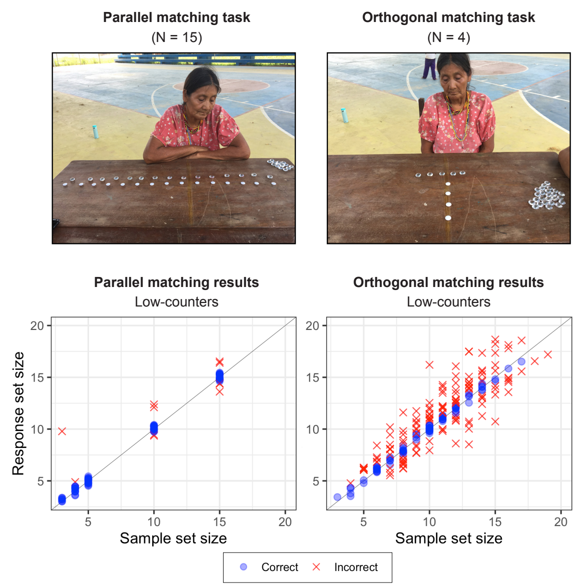

Figure 1: In the parallel matching task (top left), Tsimane' participants used 1-to-1 correspondence to make a numerical match (on sets of 3, 4, 5, 10, and 15 objects). Low-counters were highly accurate on this task (bottom left). In the orthogonal matching task (top right), correctly matching required participants to represent the cardinality of the sets (of 4-25 objects). Low-counters' accuracy was variable on this task, with signs of scalar variability (bottom right).

Results

Verbal and non-verbal number tasks

We tested participants' verbal counting abilities using a simple pebble-counting task, once before the matching tasks and again afterward (see *Materials and Methods*). Participants whose highest verbal count was 20 or less were included in the group of *low-counters* (N = 15). As a control, we also ran a group of 15 *high-counters* on the same set of tasks; these Tsimane' adults were from the same communities but had verbal counts that reached at least 40.

Participants then performed two non-verbal number tasks in which they were asked to make arrays with the same number of objects as a sample array. In the *parallel matching* task, the experimenter presented a sample array of objects (in a lateral line) for each trial and participants arranged their response array parallel to each sample array (see Figure 1, top left). Because sample and response arrays were parallel, participants could use 1-1 correspondence to perform the match in this task, spatially aligning each object in their response array with an object in the sample array without representing the cardinality of either set. For this reason, the parallel matching task does *not* test representations of large exact numbers. Rather, success on this task suggests understanding of exact numerical equivalence: for two sets to be equal in number, every element in each set must correspond to an element in the other set (38, 39). This task also functioned as a task check, ensuring that participants understood the mechanics of these numerical reproduction tasks. Participants correctly produced a parallel match for each of five sample arrays (N = 3, 4, 5, 10, and 15) and then advanced to the orthogonal matching task.

In the *orthogonal matching task*, the sample arrays were arranged sagittally in a line extending away from the participant. Participants arranged their response arrays laterally (as in the parallel matching task), in a line that was orthogonal to the sample array (see Figure 1, top right). Unlike the parallel matching task, this task precludes spatially aligning sample and response arrays, requiring participants to represent the cardinality of each set. Note that this task places minimal demands on number representations: Participants were not asked to perform any arithmetic operations and because sample and response arrays remained visible throughout each trial, they could inspect them indefinitely before finalizing their responses (which were unspeeded). In a series of practice trials, all participants correctly performed orthogonal matches for sets of size 3, 4, and 5 (with feedback) before advancing to the critical trials.

In critical trials, participants received no feedback about their performance. For high counters, the first critical trial was a sample array of 10 objects. For low-counters, the first critical trial was a sample array with two fewer objects than the participant's highest verbal count. From this starting point, we followed a pre-defined staircasing procedure (i.e. +2 for correct, -1 for incorrect) to determine the size of each sample array until participants (a) produced three incorrect response arrays for sample arrays of the same number (e.g. samples with N=15 objects), (b) correctly matched three arrays numbering 20 or more, or (c) completed 20 critical trials (see *Materials and Methods*).

Psychophysical model of numerical abilities

To evaluate the limits of participants' exact numerical representations, we analyzed their distribution of responses using a generative Bayesian data analysis (40). This model formalized a process in which participants use an "exact" system (with constant error) for smaller sets and an approximate system (with scalar variability) for larger sets. The number at which participants switched from exact to approximate representations is the participant's *switchpoint*, our dependent measure.

Formally, for the exact system (i.e., numbers below the switchpoint) we assumed that participants responded from a $Cauchy(\mu_{low} + n, \sigma_{low})$ distribution, where n is the number of objects in the sample set and μ_{low} and σ_{low} are location and scale parameters (so that $\mu_{low} \approx 0$ means responses are centered on the true value n, and $\sigma_{low} \approx 0$ means that responses cluster tightly around the mode $\mu_{low} + n$). A Cauchy distribution was used because errors in the exact system likely reflect inattention or confusion, and estimation of this distribution is robust to outliers. For the approximate system, we assumed a standard model of approximate number psychophysics (41) where subjects respond according to the distribution $Normal(n, w_i \cdot n)$, where w_i is a Weber ratio parameter that varies by individual. Putting these together, the model assumes that, when shown a sample of n objects, participant responses r follow

$$P(r \mid n, w_i, \mu_{low}, \sigma_{low}) \sim \begin{cases} Cauchy(\mu_{low} + n, \sigma_{low}) & \text{if } n \leq s_i \\ Normal(n, w_i \cdot n) & \text{if } n > s_i \end{cases} \tag{1}$$

where s_i is the switchpoint of the i'th participant. In addition, we included a hierarchical model for participant Weber ratios w_i, such that $w_i \sim Normal(\mu_W, \sigma_W)$ constrained to be positive, which means that we partially pool participant estimates of Weber fraction. We put a uniform prior on s_i between 1 and 40, a standard normal prior on μ_{low}, and $Exponential(1)$ priors on σ_{low}, μ_W, and σ_W (see *Materials and Methods*).

This model allowed us to infer the likely distribution of switchpoint values s_i from participants' pattern of behavioral responses, while accounting for the uncertainty inherent both to exact enumeration (i.e. a noise parameter for low numbers, shared across participants) and to numerical approximation (i.e. a Weber ratio fit to each participant).

All data and analysis scripts are available in the Open Science Framework repository: osf.io/me7w4/

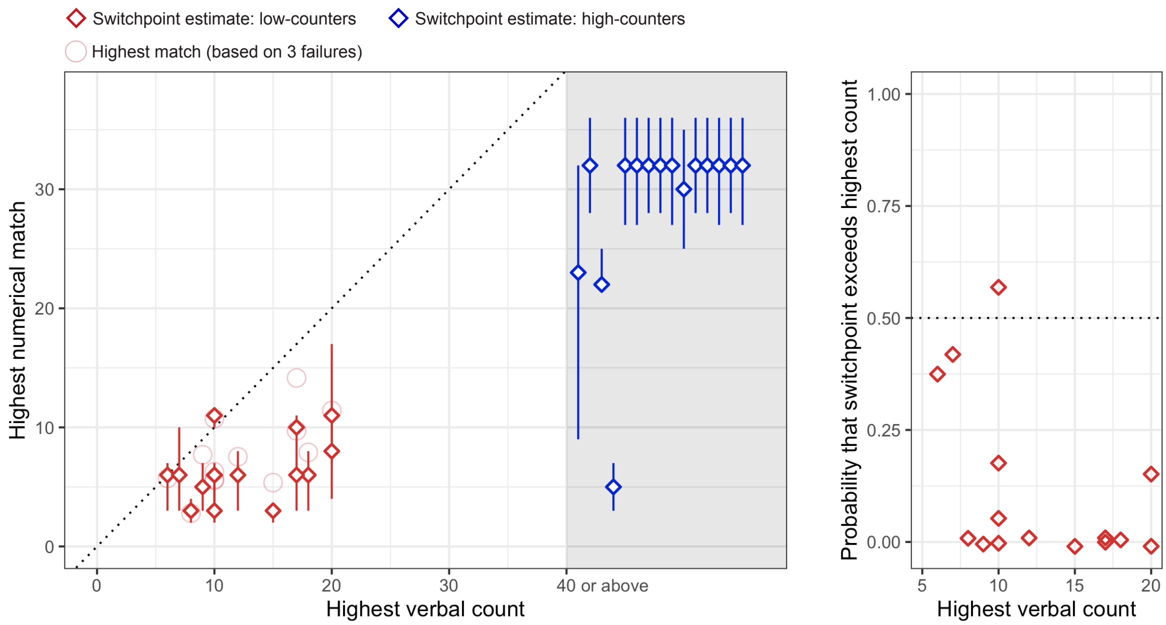

Figure 2: Left: Participants' switchpoints as a function of their highest verbal counts. Diamonds show median estimate and error bars show 50% confidence intervals. With one exception, all low-counters (red) and high counters (blue) had switchpoints *below* their highest verbal counts. The same pattern obtains for an alternative measure of participants' highest numerical match (red circles), based on the set size at which they failed three times. Right: The probability that low-counters' switchpoints exceeded their highest verbal counts.

Whereas high-counters counted to 40 without error on both trials, low-counters' highest verbal counts ranged from 6 to 20 (mean = 12.6), and often differed across the two trials (mean absolute difference = 2.0).

Parallel matching

In the parallel matching task, high-counters performed at ceiling, correctly matching each of the sample sets (i.e., N = 3, 4, 5, 10, 15) on their first attempt. Low-counters were 85% accurate on their first attempts, with 70% accuracy on sets larger than five (i.e. N = 10 and 15). With one exception, their incorrect responses were within 2 of the correct number (see Figure 1, left), and no participant made more than two errors. When they did make an error, they then showed 100% accuracy on their second attempt, fully reconstructing the response set without feedback about the magnitude or direction of their error.

Orthogonal matching

Participants were less accurate in the orthogonal matching task (mean = 51% correct) than in the parallel matching task (mean = 93% correct), even for the same cardinalities (56% correct for N = 3, 4, 5, 10, or 15; see Figure 1, top).

The model estimated a mean Weber ratio of 0.13, consistent with Weber ratios found in studies of numerical estimation in adults (26, 42), including Tsimane' adults (43). The noise for low numbers was estimated to have a mean of $\mu_{low} = -0.14$ and a standard deviation of $\sigma_{low} = 0.14$.

The critical question is how participants' switchpoints were related to their verbal counting abilities. Figure 2 (left panel) shows estimated switchpoints as a function of participants' highest verbal counts, and individual participants' data is shown in Figure 3. Although analysis of the response data was blind to participants' counting abilities, it inferred markedly different switchpoints for low and high counters, solely on the basis of their numerical matching responses. Whereas switchpoints among the low counters averaged below 7 and never exceeded 11, the average switchpoint among the high counters was over 28 ($t(17.47) = 11.01, p < .0001$). Highest verbal count reliably predicted switchpoint, above and beyond any effect of formal education: higher counters had higher switchpoints ($\beta = .55$, $SEM = 0.01$, $t = 5.48$, $p < .0001$). This relationship also held on zero education individuals: highest verbal count reliably predicted switchpoint even when analyzing only those participants with no formal education (i.e. 12 low-counters and 2 high-counters; $\beta = 0.40$, $SEM = 0.15$, $t = 2.67$, $p = .02$; all tests are two-sided).

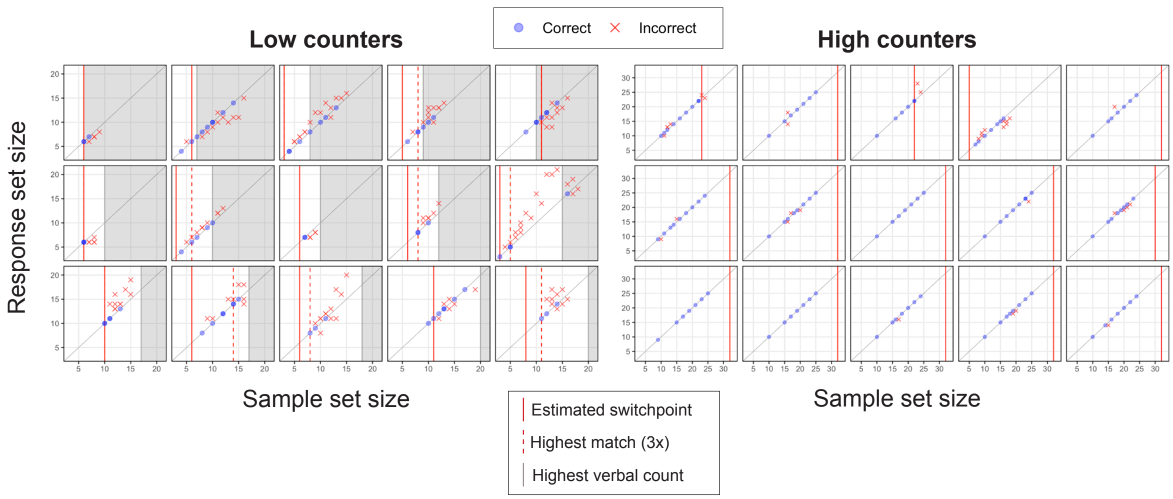

Figure 3: Each plot shows the data an individual participant. Blue dots are correct numerical matches and red Xs are incorrect responses. Shaded regions are outside the participant's verbal count range. With one exception, participants' switchpoints as estimated by the model (solid red lines) were within their verbal count range (unshaded region), as were their highest matches as determined by our 3x failure criterion (dashed red lines).

Importantly, participants' counting abilities and matching abilities were related beyond simple correlation: Low counters' switchpoints fell at or below their highest verbal count (i.e. below the diagonal dotted line) with only one exception, as shown in Figure 2 (left panel). According to Pearson chi-squared tests, this ratio (i.e. above:below) differed significantly from chance ($\chi^2(1) = 9.60$, $p = .002$). Note that in principle, participants' data points could fall below the line simply due to poor performance on the orthogonal matching task, independent of counting abilities. To assess this possibility, we conducted a permutation test in which we randomized the pairings of participants' highest verbal counts and switchpoints. This procedure respects the marginal distribution of each variable, and therefore allowed us to evaluate what proportion of data points we should expect to fall below the line by chance (i.e., if verbal counting and numerical matching performance were statistically independent). In 10,000 permuted samples, the number of participants whose switchpoints exceeded their highest counts was 7.72 on average and was never as small or smaller than the number we observed (i.e. 1), indicating that the observed pattern is extremely unlikely to occur by chance ($p < .001$).

Figure 2 (right panel) shows the probability that low counters' switchpoints exceeded their highest verbal counts, calculated using each participants' distribution of switchpoint estimates. Except for one, these switchpoints were below the 50% threshold (Mean = 11.89%), indicating that they were likely within participants' verbal count range. For numbers beyond their highest verbal count, low counters' responses were on average nearly *seven times* more likely to reflect an approximate system than an exact system.

In addition to our generative model, we also used a simple behavioral criterion to evaluate participants' highest match: the number at which they failed to produce an exact match three times (which also served as one of our stopping criteria during testing). Given the staircase procedure we used for testing, failing three times on sets of N required a combination of failing on sets of N+1 and succeeding on sets of N-2. We therefore defined highest match as two less than the number at which participants' failed three times. This alternative measure was highly correlated with participants' switchpoints as estimated by the model ($R^2 = 0.64$, $t(11) = 2.79$, $p = .02$). Although these two measures were often identical (see Figure 3), highest match was on average higher than estimated switchpoints (mean difference = 1.77), and therefore provides a more conservative estimate of participants' numerical reproduction abilities. Nevertheless, this alternative measure showed the same relationship to highest count as the switchpoint estimates from our model; with one exception, participants' highest matches were at or below their highest verbal counts (see red circles in Figure 2, left panel, and dashed red lines in Figure 3), and this ratio differed significantly from chance ($\chi^2 = 7.69$, $p = .006$). Low-counters' verbal count range reliably predicted their highest match ($t(11) = 2.44$, $p = .03$). This alternative measure also revealed the same difference between groups; whereas the highest match for low-counters (by this criteria) was below ten on average (and was always below 15), no high counter failed three times on any number we tested; rather, they all succeeded to make exact numerical matches into the twenties. (Two of the fifteen low-counters did not fail three times on the same number within 20 critical trials, and so their data do not appear in Figures 2 or 3)

Discussion

In a group of Tsimane' adults, the ability to represent exact numbers was limited to the part of the verbal count list they had mastered. Using a generative model of participants' responses, we found that they reliably matched the number of objects in a sample set only when this number was within their own verbal count range; for numbers beyond this range, participants overwhelmingly failed to make exact matches (by two measures), relying instead on numerical approximation.

Why did participants' highest matches often fall short of their highest counts, rather than equal them? In part, this gap is likely due to instability in participants' verbal count routines. Because number words are most practiced for smaller numbers, uncertainty about the next number in the list should increase as the number increases. This uncertainty results in a *soft* upper bound to the count routine, causing a given participant to falter at different numbers on different attempts. Indeed, low-counters' highest counts often differed across the two trials that we administered, which were separated by only a few minutes (mean absolute difference = 2.00). This uncertainty in highest count can also explain why one participant showed a median switchpoint slightly above their highest verbal count, as measured.

Although small gaps between highest count and highest match can be explained by the fragility of individuals' verbal count routines, larger gaps in performance may reflect a deeper conceptual limitation in some participants. A subset of our participants failed to reproduce cardinalities that were well within their count range (by as much as 12, among the low-counters), suggesting that adults with no formal education may undergo the same developmental trajectory as WEIRD children (44), who learn much of the verbal count list and the mechanics of counting long before they learn how it relates to the cardinality of sets (13, 16, 45). To understand how counting relates to cardinality, mastery of counting procedures may be *necessary* but not *sufficient*, even in adults (46).

These findings clarify the role of language in number concepts in three ways. First, unlike the Pirahã, Mundurukú, and Nicaraguan Homesigners, our Tsimane' participants succeeded in representing at least some cardinalities above four. This success shows that participants did not misunderstand the orthogonal matching task, nor were they "indifferent to exact numerical equality" (20; also see 29). On the contrary, participants were attuned to exact numerical equality in both in the orthogonal matching task and in the parallel matching task, in which they succeeded for sets as large as 15. The baseline level of numeracy we observed even among the low-counters reflects the importance of exact enumeration in the Tsimane' communities we tested, where counting practices are widespread. Yet, despite living in a numerate culture and demonstrating the ability to represent at least some large exact numbers, participants used no alternative method for representing large exact quantities. This pattern of success and failure within individuals shows that large exact number concepts are not all-or-nothing; in principle, learning part of the count list could be important for *inducing* the logic of exact enumeration, but not for representing specific cardinalities. On the contrary, we show that those representations depend critically on the availability of the corresponding (verbal) symbols.

Second, our inferences rely on comparisons across individuals, rather than across cultures or language groups (11, 25–27). Therefore, the differences in conceptual abilities that we observe cannot reflect broad differences across groups. In principle, the correlation between participants' numeric abilities could reflect differences in their formal education; on average, high-counters had more years of formal schooling (mean = 4) than low-counters (mean = 0.2). However, highest count reliably predicted highest numerical match when we controlled for differences in education, and when we analyzed only the participants with no formal schooling at all. Therefore, this relationship cannot easily be attributed to differences in language, culture, or formal education.

Finally, whereas previous studies have shown (cross-cultural) correlations between verbal counting abilities and numerical reproduction abilities, our inferences do not rely on correlation. Rather than simply asking whether one ability *predicts* the other ability, we also ask whether one ability systematically *exceeds* the other, allowing us to assess the causal relationship between them. In principle, once equipped with the logic of large exact numbers, people could represent "an unbounded set of discrete values...as needed" (22). If so, then participants' numerical matching ranges should have systematically exceeded their verbal counting ranges. We found the opposite pattern, providing the strongest evidence to date that number words play a functional role in representing large exact numbers (7, 15–18, 47).

In interpreting the findings in the Pirahã, Mundurukú, and other isolated groups, some researchers have characterized the verbal count list as a "cultural tool" (41) or a "cognitive technology" (25; also see 48). Although these metaphors may be compelling, they do little to clarify whether a verbal count list (or other external symbol system) is *necessary* for representing large exact numbers. Just as a bicycle is useful but not necessary for transportation, some scholars have argued that "using words to name exact numerosities is useful but not necessary"(23) for representing large exact numbers, providing an efficient way to encode numerical information that "complements, rather than altering or replacing, nonverbal representations" (24). If such nonverbal representations of large exact numbers exist (19, 21, 49), they had no effect on the numerical abilities of our participants (or of the Pirahã, Mundurukú, or Nicaraguan Homesigners), none of whom showed any sign of "alternative representational strategies" (24). Rather, these findings show that if the verbal count list is a cognitive technology, it is one that not only *facilitates* large exact number representations, but *enables* them.

Beyond theories of numerical cognition, these findings also bear on a broader debate about the role of language in cognition (50–55). Although *linguistic relativity* effects have been reported in a variety of domains (including color: 56, 57; time: 58; musical pitch: 59; and spatial reasoning: 60, 61), the idea that language shapes thought remains controversial (24, 62, 63), in part because there are many versions of the "Whorfian hypothesis" (64, 65). On a strong version, language can not only *change* conceptual representations but can also *enable* new ones (32, 65). The present results reveal such an effect in the domain of number, where language appears to enable representations of exact cardinalities larger than four (25–27). To be clear, language may not be the only external symbol system that can enable large exact number concepts. For example, finger counting (66), body-part counting (67), and abacus use (48) may also support the development and elaboration of such concepts (46, 68). Whatever set of symbols people use, their ability to represent large exact numbers extends only as far as their mastery of those symbols.

Materials and Methods

Participants

As part of an initial questionnaire, participants were asked to count aloud as high as they were able, starting at one, in whatever language they preferred (i.e. Tsimane' or Spanish). Those who faltered in their count routine for numbers below 20 were selected for the low-counter group and their highest count was retested using the pebble-counting task (N = 15; mean age = 48.73 +/- 4.34 years, mean schooling = 0.2 +/- .11 years) Those who initially showed good counting abilities above 40 were selected for the control group of high counters, and their highest count was retested using the pebble-counting task (N = 15; mean age = 32.87 +/- 4.52 years, mean schooling = 4.00 +/- .65 years). One participant discontinued testing before reaching any stopping criteria and was excluded from further analyses. All participants gave verbal informed consent before participating. The study was approved by the institutional review board at UC Berkeley.

Pebble counting task

Participants were given a pile of glass pebbles (N = 30 for low counters, N = 40 for high counters) on the testing table. Starting with the pebbles on their left side, participants moved them one at a time to the right while counting each one aloud. After they stopped counting, participants were asked how many pebbles there were in the counted set. The experimenter(s) and translator noted counting errors and totals given by participants. With three exceptions, participants performed this task twice, once before and once after completing the matching tasks. We used the higher of the two counts as participants' highest verbal count.

Parallel matching task

To begin the parallel matching task, an experimenter seated across from the participant laid out two white buttons (arranged left-right) and explained that the participant was to make a set of pebbles with the exact same number of objects. The experimenter then demonstrated the correct response by moving two pebbles from the participant's pile into alignment with the two buttons, making two parallel rows of two objects. Then in a series of five trials, the experimenter increased the sample array from 2 buttons to 3, 4, 5, 10, and then 15 buttons. Participants were given unlimited time to complete each match, and trials ended only after the participant verbally indicated that they had finished, at which point the response array was removed. When the response was correct, participants received verbal confirmation that their response was accurate. When participants produced a response array that differed in number from the sample, the discrepancy was pointed out and the trial was repeated. All participants completed the five trials of the parallel matching task correctly before advancing to the orthogonal matching task.

Orthogonal matching task

Like the parallel matching task, the orthogonal matching task began with a demonstration using a set size of two. The experimenter placed two buttons on the table (arranged front-back) and explained that the participant was so do the same as before: make a lateral array of pebbles with the same number of objects as the sample. In warm-up trials, participants made orthogonal matches to sets of 3, 4, and 5 buttons. If participants produced a response array in these trials that differed in number from the sample, the discrepancy was pointed out and the trial was repeated. All participants completed the three warm-up trials correctly before advancing to the critical trials. For the critical trials, low-counters began with an array of two less than their highest count and high-counters began with an array of ten. From this starting point, all participants followed the same staircasing procedure: after a correct response, the set size was increased by two; after an incorrect response, the set size was decreased by one (i.e. +2, -1).

To ensure that participants evaluated the cardinality of each sample array independently of the preceding arrays, at the end of each trial we (i) removed the response array (and reincorporated it into the larger pile of pebbles) and (ii) removed an arbitrary subset of buttons from the sample array before making the subsequent sample array. This aspect of the procedure allowed the experimenter to change the cardinality of the sample set (i.e. add two or subtract one button) out of sight of participants, making it difficult for participants to track the changes to the sample array or to infer the accuracy of their responses from those changes.

Participants were instructed to take as much time as required to ensure that their array of objects had exactly the same number of objects as in the sample array, and were free to touch the objects as needed. Each trial ended only when the participant verbally indicated that they had finished, and no feedback was given during the critical trials. The task ended when the participant (a) produced an incorrect response to three arrays of the same cardinality, (b) correctly reproduced three sets of 20 or more objects, or (c) completed 20 critical trials.

Modeling

Posterior distributions were inferred using a No-U-Turn sampler in Stan (69–71) with four chains of 10000 samples. In order to create a model with only continuous parameters, we marginalized out each participant's cutoff parameter s_i, and then computed posterior samples of those s_i from samples of other parameters. Because our behavioral responses were discrete, we computed the probability of a response r under either the Cauchy or Normal distribution as the total probability mass between $r - \frac{1}{2}$ and $r + \frac{1}{2}$. Our implementation used a non-centered parameterization of subject effects (72) and was run with $adapt\_delta = 0.9999$. With these parameters, the model encountered 209 divergent transitions in 10000 samples, but examination of a pairs plot did not reveal any regions of obvious difficulty or bias in the model. Overall, convergence was assessed by examination of the traces and computation of $\hat{R}$, which was approximately 1. The code for this model is available at osf.io/me7w4/.

Acknowledgments

This work was supported by Grant 1760874 from the National Science Foundation, Division of Research on Learning (to S.P.) and a network grant supporting B.P. from the James S. McDonnell foundation. Special thanks to Tomás Huanca, Esther Conde, and Tania Cruz for coordinating our fieldwork, to our interpreters, Manuel Roca, Robin Nate, and Elías Hiza, for their hard work, and to Kensy Cooperrider for comments on the manuscript.

Code availability

All data and analysis scripts are available in the Open Science Framework repository: osf.io/me7w4/

References

[1] Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. Core systems of number. *Trends in cognitive sciences*, 8(7):307–314, 2004.

[2] Mario Pahl, Aung Si, and Shaowu Zhang. Numerical cognition in bees and other insects. *Frontiers in psychology*, 4:162, 2013.

[3] S. Dehaene. *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, USA, 1997.

[4] Justin Halberda and Lisa Feigenson. Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology*, 44(5):1457, 2008.

[5] Stanislas Dehaene, Ghislaine Dehaene-Lambertz, and Laurent Cohen. Abstract representations of numbers in the animal and human brain. *Trends in neurosciences*, 21(8):355–361, 1998.

[6] Samuel J Cheyette and Steven T Piantadosi. A unified account of numerosity perception. *Nature Human Behaviour*, 4(12):1265–1272, 2020.

[7] Susan Carey and David Barner. Ontogenetic origins of human integer representations. *Trends in Cognitive Sciences*, 23(10):823–835, 2019.

[8] Elizabeth M Brannon. What animals know about numbers. *Handbook of mathematical cognition*, pages 85–107, 2005.

[9] Paul Bloom. Generativity within language and other cognitive domains. 1994.

[10] Noam Chomsky, Samuel Jay Keyser, et al. *Language and problems of knowledge: The Managua lectures*, volume 16. MIT press, 1988.

[11] P. Gordon. Numerical cognition without words: Evidence from Amazonia. *Science*, 306(5695):496, 2004.

[12] Karen Wynn. Children's acquisition of the number words and the counting system. *Cognitive Psychology*, 24(2):220–251, 1992.

[13] Kathryn Davidson, Kortney Eng, and David Barner. Does learning to count involve a semantic induction? *Cognition*, 123(1):162–173, 2012.

[14] Barbara W Sarnecka, Meghan C Goldman, and Emily B Slusser. How counting leads to children's first representations of exact, large numbers. *Oxford handbook of numerical cognition*, pages 291–309, 2015.

[15] Susan Carey. Bootstrapping & the origin of concepts. *Daedalus*, 133(1):59–68, 2004.

[16] Susan Carey. *The origin of concepts*. Oxford University Press, 2009.

[17] Elizabeth S Spelke. What makes us smart? core knowledge and natural language. *Language in mind: Advances in the study of language and thought*, pages 277–311, 2003.

[18] Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2):199–217, 2012.

[19] Charles R Gallistel and Rochel Gelman. Preverbal and verbal counting and computation. *Cognition*, 44(1-2):43–74, 1992.

[20] Rochel Gelman and Charles R Gallistel. Language and the origin of numerical concepts. *Science*, 306(5695):441–443, 2004.

[21] Alan M Leslie, Rochel Gelman, and CR Gallistel. The generative basis of natural number concepts. *Trends in cognitive sciences*, 12(6):213–218, 2008.

[22] Alan M Leslie, CR Gallistel, and Rochel Gelman. Where integers come from. *The innate mind: Foundations and the future*, 3:109–149, 2007.

[23] Brian Butterworth, Robert Reeve, Fiona Reynolds, and Delyth Lloyd. Numerical thought with and without words: Evidence from indigenous australian children. *Proceedings of the National Academy of Sciences*, 105(35):13179–13184, 2008.

[24] Lila Gleitman and Anna Papafragou. New perspectives on language and thought. *The Oxford handbook of thinking and reasoning*, 2:543–568, 2012.

[25] Michael C Frank, Daniel L Everett, Evelina Fedorenko, and Edward Gibson. Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, 108(3):819–824, 2008.

[26] Pierre Pica, Cathy Lemer, Véronique Izard, and Stanislas Dehaene. Exact and approximate arithmetic in an amazonian indigene group. *Science*, 306(5695):499–503, 2004.

[27] Elizabet Spaepen, Marie Coppola, Elizabeth S Spelke, Susan E Carey, and Susan Goldin-Meadow. Number without a language model. *Proceedings of the National Academy of Sciences*, 108(8): 3163–3168, 2011.

[28] Daniel Casasanto. Crying" whorf". *Science*, 307(5716):1721–1722, 2005.

[29] Stephen Laurence and Eric Margolis. Linguistic determinism and the innate basis of number. *The Innate Mind*, 3:139–169, 2007.

[30] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010.

[31] Daniel L Everett. *Don't sleep, there are snakes: Life and language in the Amazonian jungle.* Random House LLC, 2009.

[32] Rochel Gelman and Brian Butterworth. Number and language: how are they related? *Trends in cognitive sciences*, 9(1):6–10, 2005.

[33] Kensy Cooperrider and Dedre Gentner. The career of measurement. *Cognition*, 191:103942, 2019.

[34] Yoan Diekmann, Daniel Smith, Pascale Gerbault, Mark Dyble, Abigail E Page, Nikhil Chaudhary, Andrea Bamberg Migliano, and Mark G Thomas. Accurate age estimation in small-scale societies. *Proceedings of the National Academy of Sciences*, 114(31):8205–8210, 2017.

[35] Rose M Schneider and David Barner. Children use one-to-one correspondence to establish equality after learning to count. *42nd Annual Meeting of the Cognitive Science Society*, 2020.

[36] Michael C Frank, Evelina Fedorenko, Peter Lai, Rebecca Saxe, and Edward Gibson. Verbal interference suppresses exact numerical representation. *Cognitive psychology*, 64(1-2):74–92, 2012.

[37] T. Huanca. *Tsimane oral tradition, landscape, and identity in tropical forest.* SEPHIS, South-South Exchange Programme for Research on the History of Development, 2008.

[38] Julian Jara-Ettinger, Steve Piantadosi, Elizabeth S Spelke, Roger Levy, and Edward Gibson. Mastery of the logic of natural numbers is not the result of mastery of counting: Evidence from late counters. *Developmental science*, 20(6):e12459, 2017.

[39] David Hume. A treatise of human nature [1739]. *British Moralists*, 1650:1800, 1978.

[40] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[41] Stanislas Dehaene. *The number sense: How the mind creates mathematics.* OUP USA, 2011.

[42] Manuela Piazza, Véronique Izard, Philippe Pinel, Denis Le Bihan, and Stanislas Dehaene. Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44(3):547–555, 2004.

[43] Edward Gibson, Julian Jara-Ettinger, Roger Levy, and Steven Piantadosi. The use of a computer display exaggerates the connection between education and approximate number ability in remote populations. *Open Mind*, 2(1):37–46, 2017.

[44] Steven T Piantadosi, Julian Jara-Ettinger, and Edward Gibson. Children's learning of number words in an indigenous farming-foraging group. *Developmental Science*, 17(4):553–563, 2014.

[45] Karen Wynn. Children's understanding of counting. *Cognition*, 36(2):155–193, 1990.

[46] Heike Wiese. Iconic and non-iconic stages in number development: the role of language. *Trends in cognitive sciences*, 7(9):385–390, 2003.

[47] Mathieu Le Corre and Susan Carey. One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105(2):395–438, 2007.

[48] Michael C Frank and David Barner. Representing exact number visually using mental abacus. *Journal of Experimental Psychology-General*, 141(1):134, 2012.

[49] Charles R Gallistel and Rochel Gelman. Non-verbal numerical cognition: From reals to integers. *Trends in cognitive sciences*, 4(2):59–65, 2000.

[50] D. Hume. An enquiry concerning human understanding, 1748/2000.

[51] Benjamin Lee Whorf. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. Mit Press, 2012.

[52] Edward Sapir. The status of linguistics as a science. *Language*, pages 207–214, 1929.

[53] J.A. Fodor. *The language of thought*. Harvard University Press, Cambridge, MA, 1975. ISBN 0674510305.

[54] Gary Lupyan. The centrality of language in human cognition. *Language Learning*, 66(3):516–553, 2016.

[55] Stephen C Levinson and Melissa Bowerman. *Language acquisition and conceptual development*. Number 3. Cambridge University Press, 2001.

[56] Lewis Forder and Gary Lupyan. Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General*, 148(7): 1105, 2019.

[57] Terry Regier and Paul Kay. Language, thought, and color: Whorf was half right. *Trends in cognitive sciences*, 13(10):439–446, 2009.

[58] Tom Gijssels and Daniel Casasanto. Conceptualizing time in terms of space: Experimental evidence. *Cambridge handbook of cognitive linguistics*, pages 651–668, 2017.

[59] Sarah Dolscheid, Shakila Shayan, Asifa Majid, and Daniel Casasanto. The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological science*, 24(5):613–621, 2013.

[60] Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel BM Haun, and Stephen C Levinson. Can language restructure cognition? the case for space. *Trends in cognitive sciences*, 8(3):108–114, 2004.

[61] Stephen Levinson, Sérgio Meira, The Language, and Cognition Group. 'natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, pages 485–516, 2003.

[62] John H McWhorter. *The language hoax: Why the world looks the same in any language*. Oxford University Press, USA, 2014.

[63] Steven Pinker. The language instinct, william morrow and company. *Inc., New York*, 1994.

[64] Paul Kay and Willett Kempton. What is the sapir-whorf hypothesis? *American anthropologist*, 86 (1):65–79, 1984.

[65] Daniel Casasanto. Linguistic relativity. *The Routledge handbook of semantics*, pages 158–174, 2016.

[66] Andrea Bender and Sieghard Beller. Nature and culture of finger counting: Diversity and representational effects of an embodied cognitive tool. *Cognition*, 124(2):156–182, 2012.

[67] Geoffrey B Saxe. Body parts as numerals: A developmental analysis of numeration among the oksapmin in papua new guinea. *Child development*, pages 306–316, 1981.

[68] Karenleigh A Overmann. Constructing a concept of number. 2018.

[69] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[70] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37, 2016.

[71] Stan Developent Team et al. Rstan: the r interface to stan. *R package version*, 2(1), 2016.

[72] Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4, 2015.