Do Children Use Language Structure to Discover the Recursive Rules of Counting?

Rose M. Schneider,[1] Jessica Sullivan,[2] Franc Marušič,[4] Rok Žaucer,[4] Priyanka Biswas,[3] Petra Mišmaš,[4] Vesna Plesničar,[4] David Barner[1,3]

[1] Psychology Department, University of California, San Diego

[2] Department of Psychology, Skidmore College

[3] Department of Linguistics, University of California, San Diego

[4] Center for Cognitive Science of Language, University of Nova Gorica

CONTACT:

Rose M. Schneider

Psychology Department

University of California, San Diego

9500 Gilman Drive, La Jolla, CA 92093-0109

roschnei@ucsd.edu

Abstract

We test the hypothesis that children acquire the successor function — a foundational principle stating that every natural number *n* has a successor *n*+1 — by learning the productive linguistic rules that govern verbal counting. Previous studies report that speakers of languages with less complex count list morphology have greater counting and mathematical knowledge at earlier ages in comparison to speakers of more complex languages (e.g., Miller & Stigler, 1987). Here, we tested whether differences in count list transparency affected children's acquisition of the successor function in three languages with relatively transparent count lists (Cantonese, Slovenian, and English) and two languages with relatively opaque count lists (Hindi and Gujarati). We measured 3.5- to 6.5-year-old children's mastery of their count list's recursive structure with two tasks assessing productive counting, which we then related to a measure of successor function knowledge. While the more opaque languages were associated with lower counting proficiency and successor function task performance in comparison to the more transparent languages, a unique within-language analytic approach revealed a robust relationship between measures of productive counting and successor knowledge in almost every language. We conclude that learning productive rules of counting is a critical step in acquiring knowledge of recursive successor function across languages, and that the timeline for this learning varies as a function of counting transparency.

Keywords: Cross-linguistic; Count list; Successor function; Natural number concepts; Number acquisition; Conceptual development

1. Introduction

Linguistic expressions of number - like *sixteen* and *seventy-two* - provide humans with a powerful ability to exactly quantify a limitless array of objects and entities. This ability transcends our specific experience with numbers: Although most people have never counted to (or even thought about) the number *three million and thirty-two*, we can immediately recognize it as a possible number, and can judge without hesitation that adding "1" would generate *three million and thirty-three*. Somehow, when we learn to count in childhood, we use a finite training set to extract a set of rules that yield a potential infinity of numbers. In this sense, number words - like language more generally - make "infinite use of finite means" (Chomsky, 1965; von Humboldt, 1999), perhaps by virtue of the fact that they're fundamentally linguistic symbols, acquired by children as part of the process of acquiring language. In the present study, we pursued this analogy between natural language and the acquisition of counting, and tested how children's learning of recursive counting rules is affected by the grammatical structure of counting cross-linguistically. Specifically, we asked how the rule-governed structure of number word morphology in transparent languages (Cantonese, Slovenian, and English) and more opaque languages (Hindi and Gujarati) affected children's discovery of the recursive successor function that governs counting.

Although children begin reciting some portion of the count list at around 2 years of age (Fuson, 1988), they initially appear to treat it as a closed routine, or "unbreakable chain", rather than as a productive system governed by rules. At this point, many children are unable to count beyond 10, and show no understanding of how counting can be used to determine the cardinality of sets. For example, a child who is able to count to 10 may nevertheless produce a random number of items if asked to generate a set of *one* (LeCorre, Van de Walle, Brannon, & Carey, 2006; Wynn, 1990, 1992). Despite being able to recite a partial count list, children acquire meanings for the numerals *one, two,* and *three* in highly protracted stages over the course of about 18 months between the ages of 2.5 and 4 years, and rarely even attempt to count when asked to label sets or give a particular number of objects (Le Corre & Carey, 2007). Only at around the age of 3.5 to 4 years do US children begin to systematically use counting to label and generate sets, evidence that they have acquired some form of Cardinal Principle - e.g., that the final word used in a count routine labels the cardinality of the set as a whole (Gelman & Gallistel, 1978).

This apparent discontinuity in children's understanding of number is not currently well understood, though there is growing evidence that it is importantly related to changes in children's readiness for subsequent numerical learning (Geary, 2018; Geary, vanMarle, Chu, Rouder, Hoard, & Nugent, 2018; Spaepen, Gunderson, Gibson, Goldin-Meadow, & Levine, 2018). On some accounts, acquisition of the cardinal principle (CP) indicates a moment of conceptual change in which children discover the semantic content of counting (Carey, 2004, 2009; Sarnecka & Carey, 2008; LeCorre & Carey, 2007; Wynn, 1990, 1992). On this view, children make a "wild induction" based on an analogical mapping between counting and cardinality: They notice that as one counts up from *one* to *two* and then from *two* to *three*, the cardinality of the sets labeled by these words grows in increments of exactly one (see also Gentner, 2010; Marchand & Barner, 2017; Wynn; 1992). On the basis of this isomorphism between the count list and cardinal meanings, children hypothesize that the meaning of the next numeral in their list (*four*) differs from the cardinality of the previous numeral by exactly 1 as well, and that more generally every number, *n,* has a successor defined as *n*+1. This, as noted by Sarnecka and Carey (2008) amounts to acquiring implicit knowledge of the successor function, a central element of the Peano axioms, which provide a logical foundation for arithmetic, a subset of which are as follows:

1. 1 is a natural number
2. If n is a natural number, then S(n) is also a natural number
3. For every natural number n, S(n) $\neq 1$
4. If P is a property of natural numbers such that a) 1 has property P, and b) whenever a natural number has property P, so does its successor, then all natural numbers have property P, and every number has a natural successor.

Critically, this account posits that children acquire a recursive successor function at around the age of 3.5 or 4 years, allowing them to accurately label and generate sets of any size that is within their known count list. As evidence for this hypothesis, Sarnecka and Carey (2008) used a paradigm they called the "Unit Task." In this task, an experimenter placed either 4 or 5 items into a box, providing the appropriate cardinal label - e.g., "There are 4 frogs in the box." - and then added 1 or 2 additional items, while asking, "Are there 5 or 6 frogs in the box now?" Using this method, they found that only CP-knowers exhibited above-chance performance (around 66%), providing support for the view that acquisition of the successor function is related to acquisition of the CP. However, while CP-knowers as a group performed above chance on this task, many CP-knowers appeared to fail completely, raising the question of whether acquisition of the successor function is what makes children CP knowers, as Sarnecka and Carey claimed, or whether instead this knowledge is acquired sometime later.

Subsequent work has also found that although being a CP-knower may be a necessary condition for learning about the successor function (Spaepen et al., 2018), many CP-knowers lack knowledge of the successor function, and appear to master it only after becoming exceptionally strong counters (Cheung, Rubenson, & Barner, 2017; Davidson, Eng, & Barner, 2012; Wagner, Kimura, Cheung, & Barner, 2015). For example, Cheung et al. (2017) found that US children only succeed at the Unit Task for the largest numbers in their count lists by around age 5.5, and that this coincides with the moment at which they begin to claim that, rather than being finite, numbers never end (for related evidence that children first judge numbers to be infinite at this age, see Evans, 1983; Harnett & Gelman, 1998). This finding is consistent with a much earlier finding, by Secada, Fuson, and Hall. (1983), that children as old as 5 or 6 struggle with a task almost identical to the Unit Task - which they use to assess "counting-on" (i.e., the ability to add a set of 5 to a set of 1 without recounting the set of five objects after it is labeled for them).

Critical to our study, Cheung et al. (2017) note that performance on the Unit Task is best predicted by how high a child can count. Although it's perhaps unsurprising that counting experience might be related to discovering the underlying logic of the count system, it remains unknown how it might help. Nothing about memorizing a finite list guarantees that children should impute a recursive rule to this list, or conclude that numbers are infinite; after all, other lists, like the ABCs or the months of the year, aren't generated by a recursive rule. One possibility is that, as proposed by Carey and colleagues, children posit a recursive function based *purely* on a mapping between cardinalities and the ordered count list, such that the inference that numbers never end is an inductive generalization based on this analogy (Carey, 2004, 2009). However, as Cheung et al. (2017) note, another possibility is that the recursive rule takes its origin in the morpho-syntactic structure of the count list itself (see also Barner, 2017; Hurford, 1987; Rule, Dechter, & Tenenbaum, 2015; Yang, 2016). For example, a child learning English might notice that after each decade term (*twenty, thirty, forty*, etc.) the next number in the count sequence can be generated by appending the words *one* through *nine* in order. Given this, high counters' better performance on measures like the Unit Task might result not merely from more number language input in general, but instead from their knowledge of the productive morphological rules that allow numbers to be freely generated. Children's belief that numbers never end might originate in a rule that suggests that number *words* might never end.

Several studies provide preliminary evidence that young children learn the rules that govern counting prior to exhibiting errorless counting ability. First, when English-speaking children are asked to count as high as they can, their errors are non-random and often occur on decade transitions like *twenty-nine* and *thirty-nine* (Fuson, Richards, & Briars, 1982; Gould, 2017; Siegler & Robinson, 1982; Wright, 1994). If children simply memorized their count routine as an unstructured list like the alphabet, we might expect their errors to be randomly distributed. Children's specific failure to recall decade terms - but not the words preceding them - suggests that their ability to count up to these words is not driven purely by memory, but instead by the application of a rule that combines the highest known decade label with the numbers 1-9. Second, several studies have compared how children learn to count in languages which have more or less transparent morphological rules governing counting. For example, in languages like Cantonese, in which the numbers 11-99 can be generated via rule-governed combinations of the verbal labels for 1-10,(see Table 1) children count higher and make fewer errors than same-aged children learning English, which has multiple exceptions in the teens and most decade labels (Miller, Smith, Zhu, & Zhang, 1995; Miller & Stigler, 1987). Speakers of Korean, which like Cantonese is highly transparent, also exhibit better performance on multidigit addition and subtraction problems and on identifying place-value names than age-matched English-speaking children (Fuson & Kwon, 1992). In a within-culture comparison, children learning Welsh (which has a highly regular count list) outperformed English-speaking peers in tasks assessing place-value comprehension (Dowker, Bala, & Lloyd, 2008). Finally, several studies have found that children learning less transparent languages (such as English, French, and Swedish) demonstrate a weaker understanding of the base-10 system in comparison to speakers of more transparent languages like Japanese, Korean, and Cantonese (Miura, Kim, Chang, & Okamoto, 1988; Miura & Okamoto, 1989).

Although there have been several points of connection made between the regularity of counting systems, how high children can count, and mathematical achievement, no previous work has provided direct evidence that such effects are actually due to differences in how readily children extract recursive counting rules (e.g., by examining individual differences within a particular culture). For example, while previous findings are consistent with a role for counting transparency, there are also known differences in the levels of counting, number, and mathematics exposure across many of these previously studied groups (Pan, Gauvain, Liu, & Cheng, 2006; Towse & Saxton, 1998). Further, these advantages in mathematics education outcomes extend well into elementary school and high school (Siegler et al., 2012; Watts, Duncan, Siegler, & Davis-Kean, 2014) when computations depend mainly on written numerals and counting transparency should play a much weaker role, if any, in learning. Very generally, many cross-cultural differences including language, mathematics curriculum, and societal attitudes toward the importance of early math education may impact children's early counting fluency without necessarily implicating children's ability to detect recursive counting rules. If previously attested differences in mathematics learning result from the impact of counting transparency on the acquisition of counting rules, then children learning transparent counting systems should be faster to extract recursive rules from their count list, and should be faster to apply these rules to reasoning about simple addition facts, like those tested by Sarnecka and Carey's (2008) Unit Task.

The present work addresses these issues by directly assessing whether individual children have acquired productive counting rules, how knowledge of such rules is related to successor function knowledge, and how both differ across languages and cultures. First, in keeping with previous work, we tested the role of count-list transparency on acquisition of successor function knowledge by comparing children learning languages with relatively transparent count lists to children learning languages with much more opaque count lists. Second, we reasoned that if successor function knowledge is driven by learning recursive counting rules, then this should be true across all cultures, regardless of how opaque the count list, how long it takes for children to become competent counters, or how much input they receive. Third, to assess the value of this within-culture approach, we tested how other cultural factors might impact learning by comparing children learning similarly transparent languages across cultures that differ with respect to previously attested outcomes in mathematics education. To accomplish these goals, we conducted two Experiments with data from 5 different groups that varied with respect to both counting transparency and cultural practices surrounding mathematics education.

In Experiment 1, we tested children learning Cantonese (in Hong Kong), Slovenian (in Slovenia), and English (in the US). As already noted, Cantonese is a fully regular count system, with the entirety of the count list from 11-99 generated using the verbal labels for 1-10 (see Table 1). Thus, the number word *twenty-five* is formed according to the unit * decade + unit rule by concatenating the words for *two* (*yih*), *ten* (*sahp*), and *five* (*ńgh*), where *two* acts as an explicit multiplier of the base. Slovenian is slightly less transparent; while numerals from 11-100 are all formed according to a unit + unit * decade structure, it features several irregular formulations (e.g., the teens more closely resemble English). Prior to grade school, however, Slovenian children are less likely to receive counting training, and typically often cannot count past 10 when age-matched English-speaking children count to nearly 50 (Almoammer et al., 2013; Marušic, Plesničar, Razboršek, Sullivan, & Barner, 2016). Thus, while Slovenian-speaking children do have the benefit of a more regular count system, they appear to have low exposure to counting routines, allowing us to weigh the relative effects of productive counting knowledge vs. number training. Critically, however, our main question was when children learn productive counting rules, and how this knowledge is related to acquiring implicit knowledge of the successor function.

	Numeral							
	2	**5**	**10**	**20**	**25**	**50**	**52**	**102**
Cantonese	**yih**	**ńgh**	**sahp**	yihsahp	yihsahpńgh	ńghsahp	ńghsahpyih	**yātbāaklìng**yih
Slovenian	**dva**	**pet**	**deset**	**dvajset**	petindva**jset**	petdeset	dvainpetdeset	**sto** dva
English	**two**	**five**	**ten**	**twenty**	**twenty**-five	**fifty**	**fifty**-two	one **hundred** two
Gujarati	**be**	**pāṅch**	**das**	**vīs**	**panchīs**	**pachās**	**bāvan**	**eka so** be
Hindi	**do**	**pānch**	**das**	**bīs**	**pachchīs**	**pacās**	**bāvan**	**ēka sau** do

Table 1. Examples of number words in Cantonese, Slovenian, Hindi, Gujarati, and English. Languages are arranged from most to least transparent. Bolded items refer to irregular/novel items in the count list.

In Experiment 2, we investigated the relationship between recursive counting and successor function knowledge in two languages with highly opaque count lists: Hindi and Gujarati. While Hindi and Gujarati numbers are generated through a base-10 system, both exhibit many more irregularities and morphological variations in comparison to English, and these irregularities extend all the way up to 100 (Berger, 1992; Bright, 1969). For example, whereas in Hindi the numbers 2, 5, and 10 are *do*, *pānch,* and *das*, respectively, the number 20 is *bees*, 25 is *pachchīs,* and 50 is *bāvan*. Given this relatively complexity, we hypothesized that, in Hindi and Gujarati children might memorize a larger segment of their count list before converging on productive counting rules. However, as in Experiment 1, our main focus was to test the hypothesis that counting transparency matters because of its impact on the acquisition of counting rules. Therefore, we again tested within-group relations between children’s mastery of the count list’s structure and their knowledge of the successor function.

Since previous studies do not establish a single gold standard for evaluating children’s acquisition of productive counting rules, in each of these Experiments we took an exploratory approach to evaluating this knowledge. In particular, we developed and preregistered several measures of counting that sought to differentiate between children with a fully memorized list versus those who count using rules. The first was a commonly used measure of children’s counting mastery which involves testing how high they can count without error - what we call their “Initial Highest Count” (Almoammer et al., 2013; Barth, Starr, & Sullivan, 2009; Cheung et al., 2017; Davidson et al., 2012; Marušič et al., 2016; Fuson et al., 1982; Siegler & Robinson, 1982). However, a problem with this measure is that it likely underestimates children’s knowledge of productive counting rules, since many children can count higher when prompted, especially if their first error is on a decade transition, a behavior compatible with knowledge of a rule. Worse, not all children exhibit this behavior: Whereas some children - especially those who make errors on decade transitions like 29 - can count higher when prompted, other children who make errors nearby in the count list - e.g., 32 - are unable to continue counting, suggesting that their list is likely memorized, and not generated by rules (Siegler & Robinson, 1982). Given these concerns, we provided children with prompts when they made errors or omissions in their count routine. In addition to analyzing children’s initial errors when counting, we also measured their Final Highest Count, a potentially stronger indicator of productivity, and then built a measure based on the difference between these two, which we expected would provide an especially strong measure of productivity, since it reflects the ability to continue counting after an initial error when given a prompt. Finally, to test knowledge of productive rules outside of the count routine, we tested children's ability to name the next number in the count list from arbitrary points, both within and beyond the range of their Initial Highest Count (e.g., "105, what comes next?"). We reasoned that only children with strong knowledge of productive rules should perform well on this task. Taking this exploratory approach, we compared the relative ability of these metrics to predict children's acquisition of a recursive successor function as measured by Sarnecka and Carey's (2008) Unit Task, with the goal of identifying a gold-standard measure of productivity that explains knowledge across different languages and cultures.

2. Experiment 1

2.1. Method.

The methods and analyses of this study were pre-registered prior to any data collection. The pre-registration can be found at https://osf.io/tfkna/?view_only=3eb75cc3444a4187be21b152c3d5a986. All methodological and analytical choices were as pre-registered, unless stated otherwise in-text. Throughout the method section, we use English examples; however, stimuli were always presented in the child's language.

2.1.1. Participants. We pre-registered a minimum *n* of 80 per language group to conduct analyses, and a maximum *n* of 150. We recruited 378 children aged 3;6 to 6;6 from preschools, elementary schools, and the surrounding community in Hong Kong; Nova Gorica, Slovenia; and San Diego, California, USA. Fifty of these children were tested, but excluded from analyses as pre-registered for missing data from more than 20% of trials (*n* = 25); not completing the Highest Count task (*n* = 5, including children who were unable to count to *two*);[1] missing Highest Count

[1] Children who were unable to count to *two* in the Highest Count task were excluded because it was unclear whether such failure to count reflected a lack of knowledge or an unwillingness to participate.

recording (*n* = 4); being out of age range (*n* = 6); experimenter error (*n* = 6); non-native primary language (*n* = 2); noted by experimenter for exclusion (*n* = 1); or for parental interference (*n* = 1). As pre-registered, an additional two participants were excluded only from analyses involving the Next Number and WPPSI tasks in our English dataset due to failure to complete the minimum number of test trials for those tasks.

After these exclusions, our final sample included 328 participants. A breakdown of demographic information by language is shown in Table 2.

	***n* (*n* female)**	***n* CP-knower**	***M*$_{age}$ (*SD*)**	***Median*$_{age}$**
Cantonese	118 (55)	98	5.16 (0.81)	5.07
Slovenian	99 (45)	65	5.21 (0.75)	5.25
English (US)	111 (41)	71	4.79 (0.85)	4.78

Table 2. Demographic information for Cantonese, Slovenian, and US English.

2.1.2. Stimuli, design, and procedure. Children were tested individually in a room set apart from the classroom. Participants received the tasks in a fixed order (Abbreviated Give-N, Highest Count, Unit Task, Next Number, and WPPSI Picture Memory Test).

2.1.2.1. Abbreviated Give-N. This task was a conservative test of whether children understood the CP. The experimenter provided children with 10 plastic objects (e.g., buttons, bananas, apples, or bears), and a small plastic plate. After familiarizing the child with the purpose of the game, the experimenter asked them to put *N* items on the plate (trials included 6, 9, 7, and 5, in that order). After the child finished placing a set on the plate, the experimenter asked, “Is that *N?* Can you count to make sure?” If the child answered in the negative they were permitted to fix the set. If children were able to correctly generate only 3 of the 4 requested sets, they were given a second try on the failed trial. Children were classified as CP-knowers if they correctly generated sets for all four numbers. Both CP- and subset-knowers were included in analyses.

2.1.2.2. Highest Count. The experimenter introduced the task to the child by saying, “In this game I want you to count as high as you can. Can you start counting with *one*?” If the child did not begin counting after *one,* the experimenter repeated the prompt with a rising intonation. If the child made an error, the experimenter immediately stopped them by saying, “Wait a minute, what comes after *N?*” This provided the child an opportunity to self-correct. If the child failed to correct the error the experimenter provided the next number by saying, “Actually, what comes after *N* is *N+1*. Can you keep counting?” If the child did not continue, the experimenter repeated the previous 3 numbers (including the prompt) with a rising intonation to encourage the child to continue. If the child made an error immediately after being given a prompt, or was otherwise unable to continue, the experimenter stopped the task. The task was similarly ended if the child made more than 3 errors within a single decade, or more than 3 consecutive errors (i.e., counts with one prompt between each number). Otherwise, children were allowed to count to 140, and were then stopped and congratulated (“Wow! You counted to 140!”). Throughout the task, children were allowed up to 14 prompts (an average of one per decade), as we hypothesized that even children who were highly familiar with the base system might nevertheless struggle to recall decade transitions, which are often irregular. No child used all 14 prompts; the maximum number given was 12, with an average of 2.53 prompts across all languages. All children’s counting data was recorded on a voice recorder and independently coded and validated by two other researchers, which allowed us to apply the same coding criteria to all children.

2.1.2.3. Unit Task. To assess children’s understanding of the successor function, we used a modified version of the Unit Task (Sarnecka & Carey, 2008) presented on a tablet. The experimenter presented children with a scene depicting a picture of a frog on a lilypad, saying, “This is my friend, Froggie. Froggie is going to tell you about some fish she sees in the pond. You have to listen very carefully to Froggie, because she is going to ask you some questions. Let’s see what Froggie has to show us.”

Every trial had three phases. First, the child saw some number of fish move into the middle of the screen, and heard a pre-recorded female voice say in their native language, “Look! There is/are *N* fish in the pond!” The fish were visually presented for approximately 1.5s. Next, a lilypad covered the fish so that they were no longer visible, and children were given a memory check: “How many fish are in the pond?” If the child failed this first memory check, the experimenter went back to the start of the trial, saying, “Let’s try that again!” If the child failed this second memory check, the experimenter told the child how many fish were in the pond, and proceeded with the remainder of the trial.

After the memory check, children heard, “Look!” and saw one fish swim in from the right side of the screen and remain directly to the right of the lilypad. Children then heard the critical question, “Are there *N*+1 or *N*+2 fish now?” Order of alternatives (*N*+1 or *N*+2) was counterbalanced across trials. If children failed to pick one of the presented alternatives, the experimenter provided the alternatives again verbally. Participants completed a training trial (with 1 fish; on this trial, participants received feedback) and then 12 test trials (numbers queried were 5, 7, 16, 24, 52, 71, 105, 107, 116, 224, 252, and 271). In contrast to previous work using this task (Cheung et al., 2017; Davidson et al., 2012; Sarnecka & Carey, 2008), we included items so large that they that could not have been produced in the Highest Count task (224, 252, and 271).

For both the Unit Task and the Next Number task (below), the correct response for a given *N* was *N*+1. “I don’t know” responses were coded as incorrect. If a participant did not respond for a given *N*, that trial was excluded from analysis, but otherwise all numeric responses were included in analyses. Trials were additionally classified as being either within or outside of a child’s Initial Highest Count.

2.1.2.4. Next Number Task. The experimenter introduced the task by saying, “This is called ‘What Comes Next.’ In this game, I’m going to say a number, and you’ll tell me the one that comes next.” For every number, the experimenter prompted the child by saying, “*N,* what comes next?” If a child gave a response that was less than the initial prompt the experimenter reminded the child that the game was called, ‘What comes *next,*’ and allowed them to change their response. Children only received one such reminder. The numbers queried in this task were the same as in the Unit Task.

2.1.2.5. Picture Memory Task. This task was adapted for display on a tablet from the WPPSI-IV (Wecshler, 2012) picture memory task, and was included to assess children’s nonverbal working memory. Children were presented with pictures of familiar items (e.g., bell, block, and hat) and told to remember them with the prompt, “Look at this/these picture(s)!”. After either 3s (single target) or 5s (multiple targets), children saw a set containing the target and some number of distractor objects (e.g., chair, drum, and rainbow). Children were asked to touch the objects they had just seen with the prompt, “Point to the picture(s) I just showed you.” As is typical for this task, to prevent the use of verbal rehearsal strategies, children were stopped from saying the names of the objects, and were told that they had to be silent during the game.

Children received three trials with feedback at the start of the task. If they selected the incorrect items on these trials, the experimenter showed them the target items again, saying, “I showed you these pictures, so you should choose these pictures." Following WPPSI protocol, if children were younger than 4 years, they began the task with 1-item trials, while children older than 4 years began the task with 2-item trials. A response was only considered to be correct if the child correctly identified every target item. The task was terminated after 3 consecutive incorrect trials, and there were either 28 or 32 possible test trials, depending on the age of the child. As the task progressed, both the number of target items and the number of distractors increased (up to 7 targets and 12 distractors).

Children received one point for every correct trial. We summed all correct trials for each participant to obtain their raw working memory score.

2.2. Measures of productivity

2.2.1. Highest Count. This task yielded three measures of children's counting knowledge, each of which might capture knowledge of productive counting rules. Initial Highest Count was the highest number counted to without errors. As noted in the Introduction, however, this measure alone may not yield a reliable measure of productivity since the source of children's errors in a Highest Count task is often ambiguous. For example, a child who makes an error at a decade transition might do so because they have counted using a productive rule (and haven't memorized the next decade label), or because their count list is purely memorized and happens to end at a decade transition. Children often make errors and continue counting (Fuson, Richards, & Briars, 1982; Miller & Stigler, 1987), especially if errors occur on decade transitions (Siegler & Robinson, 1982) which require additional practice and memorization since decade labels are often irregular, and not predicted by a rule (Rule, et al., 2015). Thus, while Initial Highest Count yields some signal about a child's base level of counting proficiency and training with counting, it has the potential to either over- or underestimate their knowledge of productive counting rules.

Due to the ambiguity associated with Initial Highest Count, we developed two alternate measures of productive counting knowledge using the Highest Count task, both of which sought to disambiguate the source of children's counting errors. Our first alternate measure, Final Highest Count, was defined as the highest number reached during the counting task with the aid of experimenter prompts. We reasoned that this measure may better capture knowledge of counting rules, since it captures the difference between a child who, having counted to, e.g., 29, can continue counting up to, e.g., 39, when prompted with the label 30 vs. one who cannot continue counting. Our second measure was a binary classification based on the difference between these first two measures (Initial and Final Highest Count) in which we classified children as either "Resilient" or "Non-Resilient" counters, on the hypothesis that Resilient counters rely on knowledge of the base-system to count up from errors. Children were classified as Resilient if they were able to count at least 2 decades past any error (without making more than 3 errors in those two decades). This criterion allowed children 2 prompts at each decade transition within those two decades, plus one additional mid-decade or decade-beginning error. Children who were unable to meet this criterion for any error made during the Highest Count task were classified as Non-Resilient. In Experiments 1 and 2, only 9 children used 10 or more prompts, and 67 children used 5 or more prompts. Overall, Resilient and Non-Resilient counters tended to use prompts similarly, and to the same degree. This categorical classification contains some noise, however; for example, it does not distinguish between children who met these criteria after counting to 30 versus those who were able to count to 100. Additionally, it may classify children who were able to count quite high but not able to continue after a prompt as Non-Resilient. Keeping these shortcomings of the classification in mind, we report Resilience as a broad diagnostic of productive counting knowledge.

2.2.2. Next Number. The Next Number task acted as a third potential measure of counting productivity. We reasoned that children who have productive counting rules should have an easier time labeling the next number in a sequence, especially for numbers beyond their Initial Highest Count. Also, unlike the other measures, this task required children to generate the next number without the benefit of the count routine's momentum. Children's Highest Contiguous Next Number was defined as the highest number for which they were able to generate a successor, provided all previously queried items were also correct. For example, if a child responded correctly for 5, 7, and 24, but incorrectly for 16, their Highest Contiguous Next Number would be 7. For children who made an error on 1, the first item in this task, their Highest Contiguous Next Number was 0.[2]

3. Results

3.1 Highest Count

A breakdown of counting profiles by language and Resilience is shown in Table 3. Consistent with prior work, we found that Cantonese-speaking children demonstrated overall greater counting proficiency than either US or Slovenian children in both their Initial and Final Highest Counts. As expected, we also found lower levels of counting proficiency in Slovenian-speaking children in comparison to the other two languages (Table 3). Children in all three languages used experimenter prompts to a similar degree. In grouping children by Resilience, we found that Resilient counters had higher Initial and Final Highest Counts than Non-Resilient counters in all languages. Nevertheless, we still found that Cantonese-speaking children were

[2] We did not preregister a Highest Contiguous Next Number for children who failed the first trial of the Next Number task, as we did not encounter any children who failed this trial during piloting. Failure on this initial trial was rare in all datasets: *n* Cantonese = 8; *n* Slovenian = 9; *n* US English = 8; *n* Hindi = 4; *n* Indian English = 5.

able to count higher than English- and Slovenian-speaking children, regardless of Resilience classification.

Consistent with our hypothesis that acquiring productive counting rules is facilitated by increased count list transparency or exposure, we found the greatest number of Resilient counters in Cantonese, with 51% of children identified as Resilient. In contrast, only 26% of Slovenian children were classified as Resilient. Finally, 38% of children were identified as Resilient in our US sample, which has higher rates of counting exposure in comparison to Slovenian, but a lower level of count list transparency.

	n	***M* IHC (*SD*)**	***M* FHC (*SD)***	***M* Prompts (*SD)***
Cantonese				
Overall	118	73 (41.91)	94 (45.06)	2.56 (1.92)
Resilient	61	85 (40.00)	122 (25.44)	3.61 (2.19)
Non-Resilient	57	60 (40.24)	64 (42.43)	1.51 (0.63)
Slovenian				
Overall	99	27 (28.36)	44 (45.07)	2.31 (1.87)
Resilient	26	56 (38.34)	109 (36.08)	4.32 (2.46)
Non-Resilient	73	17 (13.80)	21 (16.04)	1.67 (1.06)
English (US)				
Overall	111	41 (41.53)	61 (49.19)	3.02 (2.77)
Resilient	42	64 (46.32)	110 (28.33)	5.54 (3.41)
Non-Resilient	69	27 (31.19)	31 (32.01)	1.74 (0.92)

Table 3. Counting data by language. Initial and Final Highest Count are rounded.

A visualization of counting profiles is shown in Figure 1. Consistent with our motivation for seeking alternative measures of counting productivity, we found that a majority of children (63% across all languages) who stopped before 140 were nevertheless able to count beyond their Initial Highest Count when provided with prompts. Further, many children were classified as Resilient despite having fairly low Initial Highest Counts (i.e., they could count at least 2 decades beyond their first error). While Initial Highest Count was strongly correlated with Final

Highest Count in all languages (Cantonese: *r* = .86, *p* < .0001; Slovenian: *r* = .80, *p* < .0001; English (US): *r* = .82, *p* < .0001), the frequency of children who were able to count beyond their initial errors, sometimes by many decades, indicates that this measure may not always fully capture knowledge of a productive counting rule.

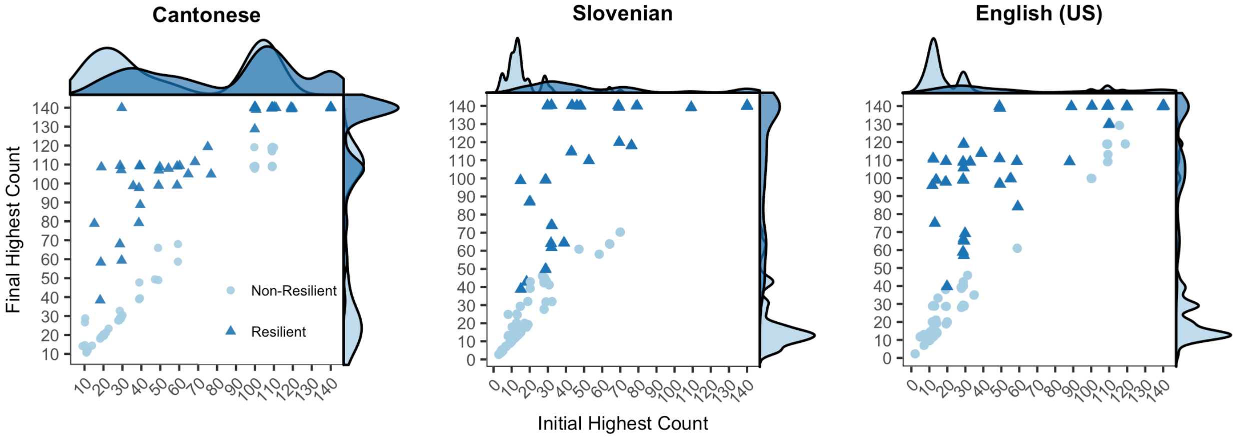

Figure 1. Initial and Final Highest Counts by language, grouped by Resilience. Points indicate the relation between a participant's Initial and Final Highest Counts. Points are jittered slightly to avoid overplotting. Density plots indicate the distribution of Initial (top) and Final (right) Highest Count by Resilience.

3.2. Predictors of successor knowledge

3.2.1. Within-language Analyses . In this section, our main goals were to test whether productive counting knowledge is predictive of successor function knowledge within each language group, and to test which measure of productivity was the strongest predictor. To address these questions, we constructed four models for each language group, separately predicting Unit Task performance from our candidate measures of productivity: (1) Initial Highest Count; (2) Final Highest Count; (3) Counting Resilience; and (4) Highest Contiguous Next Number. For each model, we also included age, whether the number was within or outside the child's Initial Highest Count, and the numerical magnitude of the Unit Task trial,[3] with

[3] Note that we pre-registered the use of numerical magnitude in these models for Experiment 2, but not for Experiment 1. Based on the results for Experiment 2, we felt that the most informative analyses were those that took into account how large the number was on the Unit Task.

subject as a random factor.[4] We tested whether each candidate measure significantly predicted Unit Task performance by conducting a Likelihood Ratio Test between each of these four individual models and the base model. Next, we constructed our large models. In constructing large models for this and all other analyses, we used hierarchical model to comparison to test whether candidate measures of productive counting knowledge explained unique or overlapping variance in children's performance. The base of each large model contained the candidate measure associated with the lowest AIC, to which we added predictors in order of increasing AIC. Candidate measures were retained on the basis of a significant χ^2 value.

	Cantonese			**Slovenian**			**English (US)**		
Predictors	*β*	*CI*	*p*	*β*	*CI*	*p*	*β*	*CI*	*p*
(Intercept)	0.69	0.44 – 0.93	<0.001	0.22	0.03 – 0.41	0.023	0.58	0.36 – 0.79	<0.001
IHC	0.81	0.56 – 1.06	<0.001	—	—	—	0.75	0.44 – 1.06	<0.001
FHC	—	—	—	0.36	0.11 – 0.61	0.004	—	—	—
HCN	—	—	—	0.35	0.12 – 0.57	0.002	0.43	0.15 – 0.71	0.003
Trial Within IHC	-0.05	-0.43 – 0.34	0.811	0.26	-0.11 – 0.64	0.173	0.09	-0.29 – 0.47	0.639
Item Magnitude	-0.52	-0.70 – -0.33	<0.001	-0.33	-0.49 – -0.17	<0.001	-0.48	-0.65 – -0.33	<0.001
Age	-0.00	-0.23 – 0.23	0.993	0.12	-0.08 – 0.31	0.241	0.03	-0.20 – 0.26	0.792

Table 4. Within-language Unit Task models in Cantonese, Slovenian, and US English. IHC = Initial Highest Count; FHC = Final Highest Count; HCNN = Highest Contiguous Next number. Only final models shown for each language: predictors without coefficient estimates did not significantly improve the fit of that language's model in a Likelihood Ratio Test. Significance was calculated using the standard normal approximation to the *t* distribution (Barr, Levy, Scheepers, & Tily, 2013).

As demonstrated in Table 4, we found that several candidate measures of counting productivity were related to acquisition of the successor function, though no single measure of productivity consistently emerged as the best predictor in each of the languages when considered

[4] Models were generalized linear mixed effects models constructed in R using the 'lme4' package (Bates, Maechler, Bolker, & Walker, 2015) with the formula: `Correct ~ [Resilience/Initial Highest Count/Final Highest Count/Highest Contiguous Next Number] + Trial Within/Outside Initial Highest Count + Item Magnitude + Age + (1|Subject)`. Continuous predictors were centered and scaled to facilitate model fit.

individually. In Cantonese, Initial Highest Count was the strongest predictor of Unit Task performance ($\chi^2_{(1)} = 38.91$, $p < .0001$), with neither Final Highest Count ($\chi^2_{(1)} = 1.39$, $p = .24$) nor Highest Contiguous Next Number ($\chi^2_{(1)} = 1.86$, $p = .17$) improving the fit of this model. In Slovenian, we found that Highest Contiguous Next Number and Final Highest Count were both predictors of Unit Task performance in Slovenian: Final Highest Count significantly improved the fit of a model containing Highest Contiguous Next Number ($\chi^2_{(1)} = 8.03$, $p = .005$), but neither the addition of Initial Highest Count ($\chi^2_{(1)} = 0.02$, $p = .89$) nor Resilience ($\chi^2_{(1)} = 0.07$, $p = .80$) explained additional variance. Finally, in English, both Initial Highest Count and Highest Contiguous Next Number were predictors of Unit Task performance: Highest Contiguous Next Number significantly improved the fit of a model containing Initial Highest Count ($\chi^2_{(1)} = 8.65$, $p = .003$), while Final Highest Count did not ($\chi^2_{(1)} = 0.10$, $p = .75$). Thus, taking this approach, we found that different measures of counting ability predicted Unit Task performance in different languages. In our next analyses, we asked which of these measures was the best fit of the entire data set (including all three language groups).

3.2.2. Cross-linguistic Analyses. In our next set of analyses, we had two goals. First, we sought to ask which of the four measures of counting ability best predicted children's Unit Task performance across all three languages when included in a single model. Second, we sought to provide a basic characterization of how the samples in our study differed, and thus whether our data are compatible with past reports that find cross-cultural differences, and in particular an advantage for Mandarin- or Cantonese-speaking children. We found that the overall pattern of performance on the Unit Task was largely consistent with prior work finding an advantage for Chinese children, with higher mean performance for Cantonese (*M* = 0.63, *SD* = 0.23) in comparison to both Slovenian (*M* = 0.55, *SD* = 0.22) and English (*M* = 0.59, *SD* = 0.25). In order to address limitations of past work, however, we attempted to account for individual differences in counting exposure by assuming that a child's Initial Highest Count offers a proxy for counting exposure. Also, unlike most previous studies, we included a working memory term to control for domain-general cognitive differences between groups. We built three models[5] predicting Unit Task performance from (1) Counting Resilience; (2) Final Highest Count; and (3) Highest Contiguous Next Number. As in our within-language analyses, these models included effects of item magnitude, whether the item was within or outside the child's Initial Highest Count, and age, with a random effect of subject. Finally, they also included children's raw working memory score.

As demonstrated in Table 5, Highest Contiguous Next Number was the single best counting productivity predictor of Unit Task performance cross-linguistically, improving the model fit in comparison to the base ($\chi^2_{(1)} = 22.70, p < .0001$). This model also revealed a significant main effect of language: when controlling for other factors, English-speaking children exhibited better performance relative to Cantonese-speaking children ($\beta = .43, p = .002$), and there were no other significant language-wise differences (Slovenian vs. Cantonese: $p = .22$; English vs. Slovenian: $p = .15$). In addition to Highest Contiguous Next Number and language, Initial Highest Count also predicted Unit Task performance. Specifically, higher Initial Highest Counts were significantly associated with better performance on the Unit Task ($\beta = 0.55, p < .001$), suggesting that, independent of knowledge of productive counting rules, exposure to the count routine is also predictive of children's successor knowledge. Critical to our hypotheses, it is important to note that these effects of Highest Contiguous Next Number, language, and

[5] Models were generalized linear mixed effects models with the formula: `Correct ~ [Resilience/Final Highest Count/Highest Contiguous Next Number] + Language*Initial Highest Count + Trial Within/Outside Initial Highest Count + Item Magnitude + Age + WPPSI + (1|Subject).` Continuous predictors were scaled and centered to facilitate model fit.

counting exposure were significant when accounting for the effects of trial difficulty, age, and working memory. Thus, the results of our model yield a more nuanced picture of the relationship between count list morphology and numerical knowledge than a comparison of mean performance alone. These data suggest that when controlling for differences in age, working memory, and exposure to the count routine, the mean differences in Unit Task performance across languages may not be best explained by count list transparency, at least not in languages with relatively small differences in count list structure, like English, Cantonese, and Slovenian.

	Comparison to Cantonese			**Comparison to Slovenian**		
Predictors	*β*	*CI*	*p*	*β*	*CI*	*p*
(Intercept)	0.31	0.10 – 0.51	0.003	0.50	0.25 – 0.76	<0.001
HCNN	0.34	0.20 – 0.47	<0.001	0.34	0.20 – 0.47	<0.001
Cantonese	—	—	—	-0.20	-0.51 – 0.12	0.225
Slovenian	0.19	-0.12 – 0.51	0.229	—	—	—
English (US)	0.43	0.16 – 0.70	0.002	0.23	-0.08 – 0.54	0.151
IHC	0.54	0.34 – 0.75	<0.001	0.55	0.24 – 0.86	0.001
Trial Within IHC	0.10	-0.12 – 0.32	0.388	0.09	-0.12 – 0.31	0.401
Item Magnitude	-0.44	-0.54 – -0.34	<0.001	-0.44	-0.54 – -0.34	<0.001
Age	0.10	-0.06 – 0.25	0.217	0.10	-0.06 – 0.25	0.231
WPPSI	0.07	-0.04 – 0.17	0.217	0.07	-0.04 – 0.18	0.219
Cantonese: IHC	—	—	—	-0.01	-0.34 – 0.32	0.966
Slovenian: IHC	0.01	-0.32 – 0.34	0.966	—	—	—
English (US): IHC	0.20	-0.08 – 0.47	0.162	0.19	-0.16 – 0.55	0.286

Table 5. Cross-linguistic Unit Task regression models with Cantonese (left) and Slovenian (right) selected as a reference group. IHC = Initial Highest Count; HCNN = Highest Contiguous Next Number.

3.3. Predictors of Next Number performance

3.3.1. Within language Analyses. In the preceding analyses, we found that Highest Contiguous Next Number emerged as the best overall predictor of successor knowledge in a model including all three languages (Cantonese, Slovenian, and English). Further, performance on the Next Number task was significantly related to Unit Task performance in all three languages when analyzed individually,[6] and was one of the strongest predictors in English and Slovenian. These findings provide support for the proposal that Next Number knowledge is a critical precursor to acquiring the successor function (Barner, 2017; Cheung et al., 2017; Davidson, et al., 2012), and also that learning the recursive rules by which number words are productively generated may support the induction that every natural number has a successor. In our next of analyses, we explored the Next Number task, testing which candidate measure of counting productivity best predicts children's performance. To do this, we evaluated the relation between children's performance on the Next Number task and Highest Count task. These analyses closely mirror our Unit Task analyses: Within each language, we constructed three models[7] predicting Next Number performance from (1) Counting Resilience; (2) Final Highest Count; and (3) Initial Highest Count.

	Cantonese			**Slovenian**			**English (US)**		
Predictors	*β*	*CI*	*p*	*β*	*CI*	*p*	*β*	*CI*	*p*
(Intercept)	-0.47	-0.78 – -0.15	0.004	-0.31	-0.62 – -0.01	0.043	-0.42	-0.75 – -0.09	0.011
IHC	1.62	1.26 – 1.98	<0.001	—	—	—	—	—	—
FHC	—	—	—	1.83	1.43 – 2.22	<0.001	1.49	1.06 – 1.92	<0.001
Trial Within IHC	0.75	0.26 – 1.24	0.003	1.19	0.69 – 1.70	<0.001	1.64	1.15 – 2.13	<0.001
Item Magnitude	-1.12	-1.40 – -0.84	<0.001	-1.09	-1.37 – -0.81	<0.001	-0.66	-0.89 – -0.42	<0.001
Age	0.34	0.01 – 0.67	0.041	0.47	0.15 – 0.80	0.004	0.69	0.27 – 1.11	0.001

Table 6. Within-language Next Number models for Cantonese, Slovenian, and US English. IHC = Initial Highest Count; FHC = Final Highest Count. Only final models shown for each language: predictors without coefficient estimates did not significantly improve the fit of that language's model in a Likelihood Ratio Test.

[6] Highest Contiguous Next Number significantly improved the fit of the Cantonese Unit Task base model ($\chi^2_{(1)}$ = 11.83, p = .0006), but did not explain unique variance when included with a model containing Initial Highest Count ($\chi^2_{(1)}$ = 1.86, p = .17).

[7] Models were generalized linear mixed effects models with the formula: `Correct ~ [Resilience/Initial Highest Count/Final Highest Count] + Trial Within/Outside Initial Highest Count + Item Magnitude + Age + (1|Subject)`. Continuous predictors were scaled and centered to facilitate model fit.

These within-language models revealed that counting ability significantly predicted Next Number performance in all three languages, although once again the best predictor differed across languages (Table 6). Final Highest Count was the best predictor of Next Number performance in Slovenian ($\chi^2_{(1)} = 72.93$, $p < .0001$) and English ($\chi^2_{(1)} = 44.93$, $p < .0001$), and Initial Highest Count did not significantly improve the fit of these models (*p*s > .05). In Cantonese, however, Initial Highest Count was the strongest predictor of Next Number performance ($\chi^2_{(1)} = 68.63$, $p < .0001$), and Final Highest Count did not explain additional variance ($\chi^2_{(1)} = 0.37$, $p = .54$). In our next set of analyses, we tested which measure best predicted Next Number performance for the entire dataset.

3.3.2. Cross-linguistic Analyses. Perhaps surprisingly, descriptive statistics indicated no evidence of an advantage for Cantonese-speakers in this task. Mean performance was around 50% for Cantonese (*M* = 0.49, *SD* = 0.35), Slovenian (*M* = 0.46, *SD* = 0.32), and English (*M* = 0.49, *SD* = 0.35). As above, we constructed models that predicted Next Number performance across languages: One predicting Next Number performance from Counting Resilience, and another predicting it from Final Highest Count. Importantly, these models tested whether these predictors remained significant when controlling for differences in overall exposure to counting by including an interaction between language and Initial Highest Count. These models also included effects of item magnitude, whether the item was within or outside the child's Initial Highest Count, and age, with a random effect of subject. Finally, these models also included a term for children's raw working memory score.

	Comparison to Cantonese			**Comparison to Slovenian**		
Predictors	*β*	*CI*	*p*	*β*	*CI*	*p*
(Intercept)	-1.51	-1.84 – -1.17	<0.001	0.27	-0.16 – 0.69	0.218
FHC	1.05	0.73 – 1.38	<0.001	1.05	0.73 – 1.38	<0.001
Cantonese	—	—	—	-1.77	-2.31 – -1.24	<0.001

Slovenian	1.77	1.24 – 2.31	<0.001	—	—	—
English (US)	1.82	1.38 – 2.25	<0.001	0.04	-0.49 – 0.57	0.881
IHC	0.65	0.28 – 1.02	0.001	0.88	0.28 – 1.48	0.004
Trial Within IHC	1.18	0.89 – 1.46	<0.001	1.18	0.89 – 1.46	<0.001
Item Magnitude	-0.92	-1.07 – -0.77	<0.001	-0.92	-1.07 – -0.77	<0.001
Age	0.61	0.34 – 0.88	<0.001	0.61	0.34 – 0.88	<0.001
WPPSI	0.15	-0.02 – 0.33	0.088	0.15	-0.02 – 0.33	0.088
Cantonese: IHC	—	—	—	-0.23	-0.81 – 0.34	0.428
Slovenian: IHC	0.23	-0.34 – 0.81	0.428	—	—	—
English (US): IHC	0.13	-0.31 – 0.56	0.571	-0.11	-0.71 – 0.49	0.730

Table 7. Cross-linguistic Next Number regression models with Cantonese (left) and Slovenian (right) selected as a reference group. FHC = Final Highest Count; IHC = Initial Highest Count.

As demonstrated in Table 7, Final Highest Count emerged as the strongest predictor of Next Number performance, and explained significant additional variance in comparison to our base model ($\chi^2_{(1)} = 38.96$, $p < .0001$); the addition of Resilience did not improve the fit of this model ($\chi^2_{(1)} = 0.27$, $p = 0.60$). Also, similar to our Unit Task models we found the surprising result that, when accounting for working memory and Initial Highest Count - a measure used by previous studies as a proxy for training with counting - Cantonese-speaking children's performance was actually significantly poorer than that of both English- ($\beta = -1.82$, $p < .001$) and Slovenian-speaking children ($\beta = 11.77$, $p < .001$), while there was no difference in performance between English and Slovenian children ($\beta = 0.04$, $p = .88$, Figure 2).

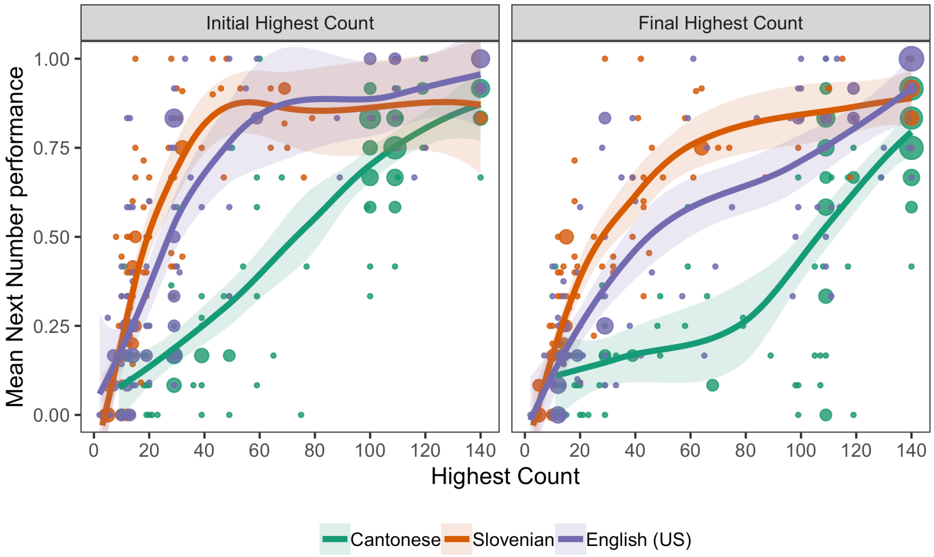

Figure 2. Scatterplot relating Highest Count (Initial, left and Final, right) to mean Next Number performance by Language. Smooth curve fitted by locally weighted regression, and shaded areas indicate 95% confidence intervals. Size of points indicate frequency of highest count.

4. Discussion

In three languages with relatively transparent counting systems, we investigated the relation between knowledge of counting structure and acquisition of the successor function. Consistent with previous reports, we found striking cross-cultural differences in counting proficiency and numerical knowledge that were broadly related to language transparency; Cantonese-speaking children were able to count much higher than English- and Slovenian-speaking children, and also had greater mean successor task performance. Despite these cross-linguistic differences, however, our within-language analyses in Cantonese, Slovenian, and English each found significant relations between measures of counting productivity and performance on the Unit Task, our measure of successor function knowledge. Also, models including all three languages found that performance on the Next Number task, a measure of children's ability to count-up from arbitrary points in the count list, was the best overall predictor of performance on the Unit Task. Additionally, we found that while Initial Highest Count predicted performance on both the Unit and Next Number tasks in all three languages, it was never *solely* predictive, and other productivity measures such as Next Number performance or Final Highest Count were better indicators of children's performance on these two tasks in English and Slovenian. In particular, the relationship between children's Next Number and Unit Task performance suggests that successor function knowledge arises in part from knowledge of productive base-10 rules.

A key element of our design was to assess the role of language structure on number knowledge within languages, rather than relying on cross-linguistic comparisons. The rationale for this was a concern that children learning different languages also typically belong to different cultures, with different practices surrounding early numeracy. Children learning Cantonese, for example, may outperform children learning English not because of the relatively small differences between their structures, but because they receive more intensive math training in early elementary school years (Towse & Saxton, 1998). This concern was vindicated by our analyses that included all three language groups. Although Cantonese-speaking children performed better than other groups on most measures, when factors like working memory, age, and a child's Initial Highest Count (a reasonable proxy for training exposure) were included in models, they were either not different from other groups, or performed significantly worse.

Still, as already noted, the grammatical differences in counting transparency between Cantonese, English, and Slovenian are relatively modest. For the most part, English and Slovenian have relatively transparent decade rules that recur from the 20's through to 100, with only decade labels and syntax presenting points of contrast with Cantonese. Other languages, however, like Hindi and Gujarati, are substantially less transparent, and feature multiple exceptions in every decade all the way to 100. We explore these languages in Experiment 2.

5. Experiment 2

In Experiment 1, we contrasted three languages which, though different with respect to base-system transparency, were nevertheless relatively similar in structure, and generally transparent in nature. In Experiment 2, we investigated two languages that are substantially less transparent in nature: Hindi and Gujarati, as well as a sample of English-speaking Indian children living in the same region of India. Based on Experiment 1, we predicted that in these languages counting productivity would again be predictive of Unit Task performance, but that far fewer children should be productive overall.

5.1. Method

As in Experiment 1, the methods and analyses of this study were pre-registered prior to any data collection. The pre-registration can be found at https://osf.io/5zxrt/?view_only=8a3165b0738748a6be074e01c18d97e7.

5.1.1. Participants. We pre-registered a minimum *n* of 80 per language group to conduct analyses, and a maximum *n* of 150. We recruited 288 children aged 3;6 to 6;6 from Hindi, Gujarati, and English medium schools in Vadodara, Gujarat, India. Children's inclusion in a particular dataset (Hindi, Gujarati, or English), was determined by their primary language of instruction. While Hindi- and Gujarati-speaking children were generally schooled in the language spoken at home, English-speaking children spoke Gujarati, Hindi, or another language at home (based on parental report).

As per our pre-registration, 47 of these children were excluded from all analyses for missing data from more than 20% of trials (*n* = 19); not completing the Highest Count task (*n* = 9); failure to comprehend the task (*n* = 8); missing Highest Count recording (*n* = 4); being out of age range (*n* = 3); experimenter error (*n* = 2); language impairment (*n* = 1); or, for Gujarati and Hindi, native language other than the language of instruction (*n* = 1). In addition to the 47 who were excluded from all analyses, 24 were removed from a subset of analyses if they did not complete the pre-registered minimum number of test trials for that task. After these exclusions, our final sample included 241 participants (Table 8).

	***n* (*n* female)**	***n* CP-knower**	***M*$_{age}$ (*SD*)**	***Median*$_{age}$**
Hindi	91 (41)	44	5.63 (0.68)	5.81
Gujarati	80 (48)	50	5.65 (0.41)	5.72
English (US)	111 (41)	71	4.79 (0.85)	4.78
English (India)	70 (35)	54	5.24 (0.85)	5.37

Table 8. Demographic information for Hindi, Gujarati, and English (US and Indian).

These exclusions had their greatest effect on the Indian English dataset, perhaps because this was the only group tested in their second language. Consequently we did not reach our minimum *n* defined in our pre-registration, and instead conducted our primary analyses, as preregistered, using the US English dataset from Experiment 1. Because US children's performance may substantially differ from that of Hindi- and Gujarati-speaking children for many reasons other than language, we interpreted these results of these analyses with caution, and compared them to *post hoc* analyses which included the Indian English dataset for tasks in which we observed low exclusion rates (Highest Count and Next Number). Thus, our Indian English dataset helped to isolate the role of language in children's performance on these measures.

5.1.2. Stimuli, methods, and procedure. These were identical to Experiment 1, with the exception that audio prompts were translated into the language of instruction.

6. Results

6.1. Highest Count

Table 9 shows a breakdown of counting profiles by language and Resilience. Overall, counting proficiency seemed to be strongly related to count list transparency; Hindi- and

Gujarati-speaking children had much lower Initial and Final Highest counts in comparison to English-speaking children. As in Experiment 1 these effects of language persisted when children were grouped by Resilience, but to a lesser degree; Resilient counters had higher Initial and Final Highest Counts than Non-Resilient counters in all three languages, although counts still tended to be higher for English-speaking children. Critically, we found that very few children were able to meet the criteria for Resilience in Hindi and Gujarati (9% and 14% of children in each language respectively). On the other hand, 40% of Indian English-speaking children were identified as Resilient, which is similar to the proportion observed in our US English sample (38%).

	n	***M* IHC (*SD*)**	***M* FHC (*SD)***	***M* Prompts (*SD)***
Hindi				
Overall	91	24 (13.63)	31 (22.99)	2.00 (1.37)
Resilient	8	38 (22.04)	81 (43.28)	4.00 (1.85)
Non-Resilient	83	22 (11.86)	26 (12.42)	1.80 (1.14)
Gujarati				
Overall	80	27 (17.75)	37 (25.87)	2.26 (1.25)
Resilient	11	52 (27.28)	86 (31.34	3.36 (0.81)
Non-Resilient	69	23 (11.55)	29 (13.64)	2.08 (1.22)
English (India)				
Overall	70	48 (39.64)	75 (48.62)	3.07 (2.12)
Resilient	28	59 (37.58)	117 (22.60)	4.61 (2.38)
Non-Resilient	42	40 (39.57)	47 (40.15)	2.05 (1.08)

Table 9. Counting data by language. Initial and Final Highest Count are rounded.

Once again, we found that a majority of children (62% across all three languages) were able to continue counting past their Initial Highest Count if they stopped prior to 140, although very few children counted far enough past their first error to be classified as Resilient in Hindi and Gujarati (Figure 3). Thus, while Initial Highest Count was strongly correlated with Final Highest Count (English (India): $r = .77, p < .0001$; Hindi: $r = .77, p < .0001$; Gujarati: $r = .91, p < .0001$), we again found evidence that it may underestimate children’s productive counting knowledge.

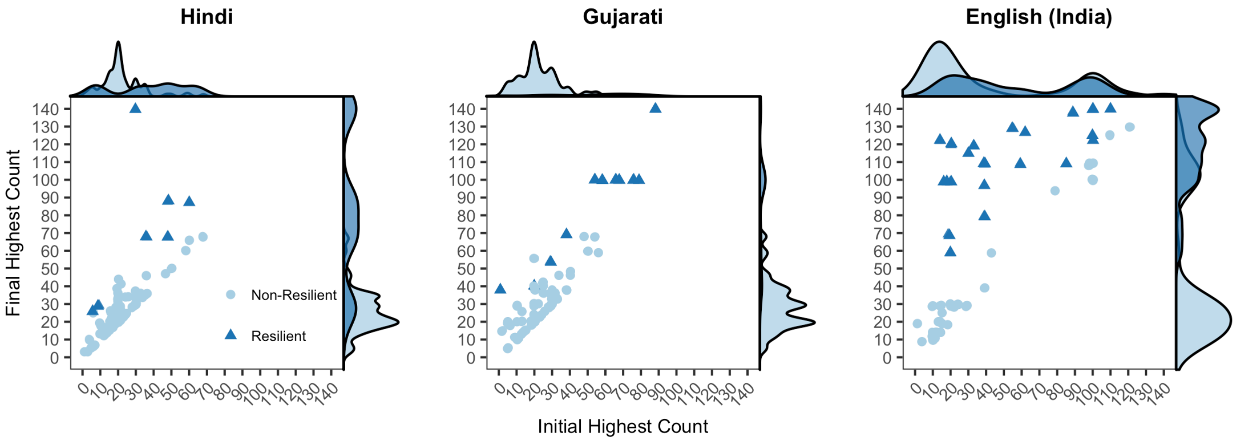

Figure 3. Initial and Final Highest Counts by language, grouped by Resilience. Points indicate the relation between a participant's Initial and Final Highest Counts. Points are jittered slightly to avoid overplotting. Density plots indicate the distribution of Initial (top) and Final (right) by Resilience.

6.2. Predictors of successor knowledge

6.2.1. Within-language Analyses. We next tested whether candidate measures of productive counting were significantly predictive of Unit Task performance in Hindi and Gujarati. We predicted that although children learning these languages may become productive counters later in development, the same relation between productive counting and successor function knowledge should nevertheless exist. The individual measures, model specifications, and hierarchical model comparisons used to assess the relationship between productive counting and successor function knowledge were identical to Experiment 1. Once again, our candidate measures of productivity were (1) Counting Resilience; (2) Final Highest Count; (3) Initial Highest Count; and (4) Highest Contiguous Next Number.

We found evidence of a link between knowledge of productive counting rules and the acquisition of the successor function in Hindi, but not in Gujarati, perhaps because so few Gujarati-speaking children exhibited knowledge of either productive rules or the successor function (see Table 10). In Hindi, Initial Highest Count and Highest Contiguous Next Number were the strongest predictors of Unit Task performance: the addition of Highest Contiguous Next

Number significantly improved the fit of a model containing only Initial Highest count ($\chi^2_{(1)} = 9.91$, $p = .002$), but the addition of Final Highest Count to a model containing both Initial Highest Count and Highest Contiguous Next Number did not ($\chi^2_{(1)} = 0.93$, $p = .34$). However, none of our candidate measures of productivity were related to successor knowledge in Gujarati, though Unit Task performance was predicted by age ($\beta = .19$, $p < .03$), and performance was better for items within a participant's Initial Highest Count range ($\beta = .54$, $p < .005$).

	Hindi			**Gujarati**			**English (US)**		
Predictors	*β*	*CI*	*p*	*β*	*CI*	*p*	*β*	*CI*	*p*
(Intercept)	-0.48	-0.67 – -0.29	<0.001	-0.35	-0.54 – -0.16	<0.001	0.58	0.36 – 0.79	<0.001
IHC	0.35	0.14 – 0.56	0.001	—	—	—	0.75	0.44 – 1.06	<0.001
FHC	—	—	—	—	—	—	—	—	—
HCNN	0.33	0.12 – 0.53	0.002	—	—	—	0.43	0.15 – 0.71	0.003
Trial Within IHC	0.03	-0.35 – 0.40	0.894	0.54	0.16 – 0.92	0.005	0.09	-0.29 – 0.47	0.639
Item Magnitude	-0.39	-0.57 – -0.22	<0.001	-0.16	-0.33 – 0.01	0.063	-0.48	-0.65 – -0.33	<0.001
Age	0.30	0.13 – 0.48	0.001	0.19	0.02 – 0.35	0.026	0.03	-0.20 – 0.26	0.792

Table 10. Within-language Unit Task models for Hindi, Gujarati, and US English**.** Only final models shown for each language: predictors without coefficient estimates did not significantly improve the fit of that language's model in a Likelihood Ratio Test. IHC = Initial Highest Count; FHC = Final Highest Count; HCNN = Highest Contiguous Next Number.

6.2.2. Cross-linguistic Analyses. In our cross-linguistic analyses we tested whether measures of counting productivity were related to Unit Task performance when all languages were included in a single model. Consistent with the hypothesis that learning a more morphologically complex count list may impede extraction of productive counting rules and the successor function, we found lower mean performance for Hindi (*M* = .40, *SD* = .22) and Gujarati (*M* = .45, *SD* = .18) children in comparison to US English (*M* = .59, *SD* = 0.25). However, as in Experiment 1, we worried that cross-cultural factors other than language might explain these differences. The following analyses explore this possibility in two steps. First, we built three separate models predicting Unit Task performance from (1) Counting Productivity,

(2) Final Highest Count, and (3) Highest Contiguous Next Number in US English, Hindi, and Gujarati. Second, we conducted *post hoc* tests that included subsets of data from children learning Indian English, who are culturally more similar to the Hindi and Gujarati children than US children, but have learned the more regular English count system. Model specifications, measures, and comparison process were identical to Experiment 1.

In the model including Hindi, Gujarati, and US English data, Highest Contiguous Next Number emerged as the strongest predictor of Unit Task performance, and significantly improved model fit compared to our base model ($\chi^2_{(1)} = 13.18$, $p = .0003$; see Table 11). In contrast to Experiment 1, however, mean differences in Unit Task performance by language persisted even when controlling for between-group differences: Hindi- and Gujarati-speaking children had significantly lower Unit Task scores compared to English-speaking US children (Hindi: $\beta = -.70.$, $p < .0001$; Gujarati: $\beta = -.72$, $p < .0001$). This much lower performance on the Unit Task for Hindi- and Gujarati-speaking children in comparison to English-speaking children suggests that acquiring the successor function may be more difficult in languages in which the recursion of the count list is less easily discoverable due to less transparent morphology. However, this conclusion is tempered by the fact that only US data were available for this particular analysis. To further probe whether this difference might be due to language, in particular, our next analyses, which focused on the Next Number task, included *post hoc* tests with Indian English data.

	Comparison to English (US)		
Predictors	*B*	*CI*	*p*
(Intercept)	0.43	0.23 – 0.63	<0.001
HCNN	0.26	0.12 – 0.40	<0.001
Hindi	-0.70	-1.02 – -0.39	<0.001
Gujarati	-0.72	-1.02 – -0.43	<0.001

IHC	0.45	0.27 – 0.63	<0.001
Within IHC	0.16	-0.06 – 0.39	0.148
Item Magnitude	-0.37	-0.47 – -0.27	<0.001
Age	0.27	0.11 – 0.44	0.001
WPPSI	0.10	-0.01 – 0.21	0.071
Hindi:IHC	0.45	0.04 – 0.87	0.032
Gujarati:IHC	-0.42	-0.75 – -0.09	0.012

Table 11. Cross-linguistic Unit Task models with US English as a reference group. HCNN = Highest Contiguous Next Number; IHC = Initial Highest Count.

6.3. Predictors of Next Number Performance

6.3.1. Within-language Analyses. The results of our Unit Task analyses indicated that although productive counting is related to successor knowledge in some children learning opaque count lists, this connection is somewhat more fragile. One potential factor limiting our ability to detect effects was the relative infrequency of children who demonstrated productive counting knowledge in these languages: In Hindi only 8 out of 91 children were classified as Resilient, and in Gujarati, only 11 out of 80. Thus, we surely lacked power to reliably detect relations between productivity and other outcomes. Nevertheless, in both Hindi and our cross-linguistic Unit Task analyses, Highest Contiguous Next Number again significantly predicted successor knowledge. As in Experiment 1, we next explored the best predictors of Next Number performance. We again constructed three within-language models as in Experiment 1, predicting Next Number performance from (1) Counting Resilience; (2) Final Highest Count; and (3) Initial Highest Count. Because we observed higher rates of comprehension in our Indian English sample for this task relative to other tasks, we include the results of their within-language analyses here.

	Hindi			**Gujarati**			**English (India)**		
Predictors	*β*	*CI*	*p*	*β*	*CI*	*p*	*β*	*CI*	*p*
(Intercept)	-1.88	-2.34 – -1.43	<0.001	-1.24	-1.69 – -0.78	<0.001	-0.76	-1.13 – -0.39	<0.001
IHC	1.41	0.93 – 1.88	<0.001	—	—	—	0.64	0.13 – 1.15	0.014

FHC	__	__	__	1.78	1.13 – 2.42	<0.001	0.85	0.25 – 1.44	0.005
Resilient	__	__	__	-1.75	-3.46 –-0.04	0.045	—	—	—
Trial Within IHC	1.68	1.14 – 2.22	<0.001	2.00	1.43 – 2.57	<0.001	1.11	0.52 – 1.70	<0.001
Item Magnitude	-0.92	-1.24 – -0.61	<0.001	-1.25	-1.59 – -0.91	<0.001	-1.37	-1.72 – -1.02	<0.001
Age	0.53	0.09 – 0.98	0.020	0.37	-0.02 – 0.75	0.064	0.18	-0.30 – 0.67	0.458

Table 12. Within-language Next Number models for Hindi, Gujarati, and Indian English. IHC = Initial Highest Count; FHC = Final Highest Count. Only final models shown for each language: predictors without coefficient estimates did not significantly improve the fit of that language's model in a Likelihood Ratio Test.

Despite the lower levels of counting proficiency in Hindi and Gujarati, we again found that counting ability was significantly predictive of Next Number performance in all languages (see Table 12). In Hindi, although Initial Highest Count, Final Highest Count, and Resilience were each significantly related to Next Number performance, overall Initial Highest Count was the strongest predictor ($\chi^2_{(1)} = 33.412$, $p < .0001$), and neither Final Highest Count ($\chi^2_{(1)} = 2.77$, $p = .10$) nor Resilience ($\chi^2_{(1)} = 0.37$, $p = .54$) significantly improved model fit. In Gujarati, Final Highest Count and Resilience were the best predictors of performance; Resilience significantly improved the fit of a model containing Final Highest Count ($\chi^2_{(1)} = 4.12$, $p = .04$), and Initial Highest Count did not add to this model ($\chi^2_{(1)} = .08$, $p = .78$). Finally, in Indian English, both Initial and Final Highest count produced the best fit to the data in comparison to a model containing Final Highest Count alone. ($\chi^2_{(1)} = 5.88$, $p = .02$).

6.3.2. Cross-linguistic Analyses. Although we found that counting ability predicted Next Number performance in our within-language analyses, mean performance on the Next Number task was lower in Hindi (*M* = .32, *SD* = .30) and Gujarati (*M* = .38, *SD* = .26) in comparison to both US English (*M* = .49, *SD* = .35) and Indian English (*M* = .45, *SD* = .29), which did not significantly differ from one another (*t*(177) = 0.88, *p* = .38). As pre-registered, we constructed cross-linguistic models using our US English dataset. Because we observed a much higher rate of comprehension for the Next Number task in our Indian English sample, however, we also report *post hoc* analyses including these data in an attempt to isolate the effects of language transparency versus other cultural factors. In these analyses, we again controlled for between-group differences by including an interaction between language and Initial Highest Count, as well as a nonverbal working memory term. Once again, the model specifications and comparison process were identical to Experiment 1. Using this base model we constructed two generalized linear mixed effects models predicting Next Number performance from (1) Counting Resilience; and (2) Final Highest Count.

	Comparison to English (US)			**Comparison to English (India)**		
Predictors	*β*	*CI*	*p*	*β*	*CI*	*p*
(Intercept)	-0.63	-1.06 – -0.19	0.005	-1.33	-1.88 – -0.77	<0.001
FHC	0.98	0.58 – 1.38	<0.001	0.66	0.21 – 1.11	0.004
Hindi	-0.53	-1.24 – 0.17	0.136	0.30	-0.46 – 1.07	0.438
Gujarati	-0.53	-1.18 – 0.12	0.112	0.29	-0.42 – 1.01	0.422
IHC	0.42	0.00 – 0.84	0.047	0.25	-0.18 – 0.68	0.262
Trial Within IHC	1.73	1.42 – 2.04	<0.001	1.66	1.33 – 2.00	<0.001
Item Magnitude	-0.91	-1.07 – -0.74	<0.001	-1.21	-1.41 – -1.01	<0.001
Age	0.80	0.43 – 1.16	<0.001	0.57	0.16 – 0.99	0.006
WPPSI	0.15	-0.08 – 0.37	0.210	0.13	-0.11 – 0.37	0.293
Hindi:IHC	1.48	0.58 – 2.38	0.001	2.05	1.19 – 2.90	<0.001
Gujarati:IHC	0.45	-0.28 – 1.18	0.227	1.00	0.29 – 1.71	0.006

Table 13. Cross-linguistic Next Number models with US English (left) and Indian English (right) selected as a reference group. FHC = Final Highest Count; IHC = Initial Highest Count.

As shown in Table 13, Final Highest Count was the strongest single predictor of Next Number performance in Hindi, Gujarati, and US English, significantly improving the fit of the base model ($\chi^2_{(1)} = 22.68$, $p < .0001$). A follow-up analysis which substituted our Indian for US English dataset as the comparison group replicated this effect ($\chi^2_{(1)} = 8.36$, $p = .004$). There was no effect of language in either the US or Indian English models (Figure 4), suggesting that although counting mastery may vary across languages due to morphological complexity, the best predictor of performance on the Next Number task is nevertheless children's knowledge of productive counting rules, reflected here by their Final Highest Count.

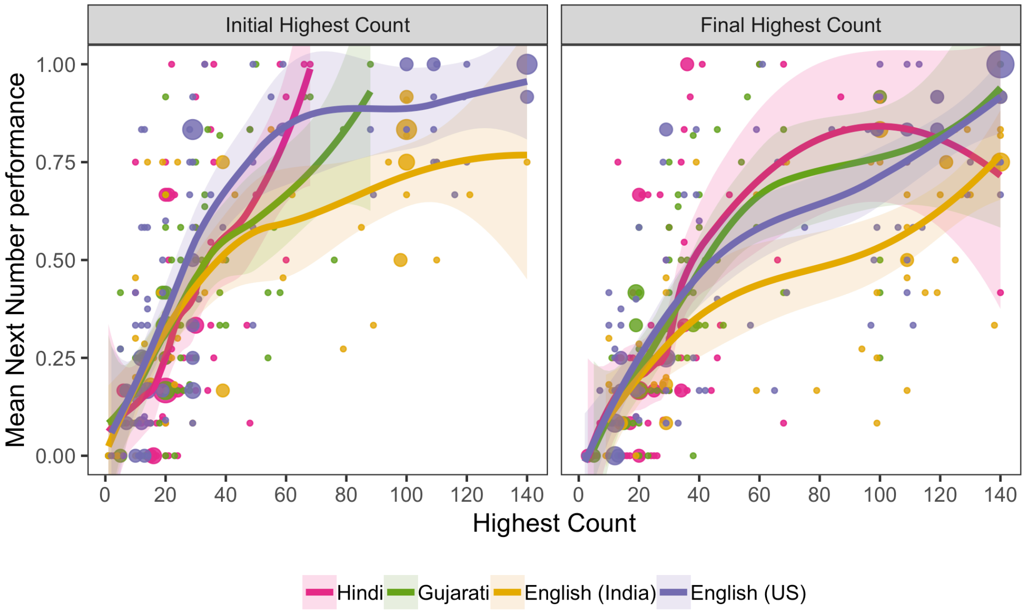

Figure 4. Scatterplot relating Highest Count (Initial, left and Final, right) to mean Next Number performance by language. Smooth curve fitted by locally weighted regression, and shaded areas indicate 95% confidence intervals. Size of points indicate frequency of highest count.

7. Discussion

In Experiment 2, we investigated children learning Hindi and Gujarati, two languages that are substantially less transparent in nature than any of the languages studied in Experiment 1, to explore whether acquiring the successor function was made more difficult by a more complex count list. We again found substantial differences in counting ability and successor knowledge between transparent and opaque language groups, and these differences were greater than those reported in Experiment 1, even in a within-culture comparison. Using a within-language approach, however, we found that productive counting knowledge was related to Unit Task performance, despite lower numbers of productive counters in Hindi and Gujarati. Although this evidence was less robust than in Experiment 1, likely due to overall lower levels of counting ability in Hindi and Gujarati, we again found that knowledge of productive counting was significantly related to successor knowledge in Hindi. While the best predictors of Unit Task performance differed across our within-language models, once again our combined cross-linguistic models revealed Next Number performance, an indicator of children’s ability to count up from an arbitrary point in the count list, as the strongest indicator of children’s successor knowledge. Similarly, although there was some variability in the counting measures related to Next Number performance in within-language models, Final Highest Count emerged as the best predictor on this task in our cross-linguistic models. These results mirror our findings from Experiment 1; once again, we find that two strong indicators of children’s base-10 knowledge are significantly related to their performance on the Unit and Next Number tasks.

In contrast to Experiment 1, however, our cross-linguistic models also revealed strong effects of count list transparency. Children learning less complex count lists seemed to benefit from this transparency in acquiring the successor function, even when controlling for factors such as age, working memory, and individual differences in Initial Highest Count; English-speaking US children performed better on the Unit Task relative to Hindi and Gujarati children. Thus, unlike in Experiment 1 where mean cross-linguistic differences disappeared when accounting for other factors, the significantly lower performance of Hindi and Gujarati children in comparison to English-speaking US children on the Unit Task suggests that extremely opaque count lists may confer a substantial disadvantage in acquiring other numerical knowledge. While these findings are tempered by the fact that we were unable to make a within-culture comparison for the Unit Task using our Indian English sample, *post hoc* comparisons indicated that these children’s performance on the Highest Count and Next Number tasks was comparable to US English children, despite being tested in their second language.

8. General Discussion

Using a large cross-linguistic dataset drawn from five languages across four cultures, we tested the hypothesis that children acquire the successor function through learning the productive morphological rules of their language's count list, while also exploring which candidate measures of productivity best predict successor function knowledge. We found large mean differences in counting ability and successor knowledge between languages with relatively transparent counting systems (Cantonese, Slovenian, English) compared to those with more opaque counting structure (Hindi, Gujarati). Also, despite these cross-linguistic differences, we found that productive counting knowledge was strongly related to successor function knowledge within each language.

In addition to these main findings, our study also revealed several important secondary results. First, although measures of counting productivity were related to successor function knowledge across languages, there was important variability in which measures of counting productivity best predicted successor function knowledge. Second, despite these interesting differences, when all languages were considered together, children's successor function knowledge in both Experiments 1 and 2 was best predicted by performance on the Next Number task - i.e., the ability to name the next number in the count list without counting up from 1. Third, somewhat surprisingly, we found that although, like in past reports, Cantonese-speaking children outperformed English-speaking children along several measures of number knowledge, these differences disappeared or were reversed when other factors were considered, like working memory and amount of counting exposure. Similarly, although Slovenian is more transparent than English, Slovenian children exhibited much more limited counting abilities than English-speaking children. These results lead us to conclude that some previously reported cross-linguistic differences may be not be due to language alone, but may also be importantly affected by cultural differences in math practices.

This work began with two observations. First, previous work reported that children who learn relatively transparent languages, like Cantonese, make fewer counting errors than children learning English, and may be quicker to acquire early mathematical abilities. Second, previous studies found that children's acquisition of the successor function, as measured by the Unit Task, is strongly predicted by how high they can count. Together, these two observations led us to hypothesize that exposure to counting might lead children to move beyond a memorized list to derive productive rules that allow them to count indefinitely high, and that these rules might be the basis for deriving a recursive successor function: Knowledge that counting *words* are governed by a rule-like structure might lead children to infer that numbers themselves are generated by a recursive '+1' rule. To investigate this question, we took a novel approach. Although past work has found connections between the regularity of counting systems, counting proficiency, and mathematical achievement, none of this work has provided direct evidence that such effects are actually due to differences in how easily children are able to extract productive counting rules through exploring individual differences within a particular culture. For example, although previous findings (Fuson & Kwon, 1992; Miller, Kelly, & Zhou, 2005; Miller & Stigler, 1987; Miura et al., 1988; Miura & Okamoto, 1989) are consistent with the idea that such advantages are due to count list transparency, it is possible that they may also reflect differences in the levels of counting, number, and mathematics exposure across these groups (Pan et al., 2006; Towse & Saxton, 1998). Very generally, many cross-cultural differences including language, mathematics curriculum, and societal attitudes toward the importance of early math education may impact children's early counting fluency without necessarily implicating children's ability to detect recursive counting rules. If previously attested differences in mathematics learning result from the impact of counting transparency on the acquisition of counting rules, then children learning transparent counting systems should be faster to extract recursive rules from their count list, and should be faster to apply these rules to reasoning about simple addition facts, like those tested by Sarnecka and Carey's (2008) Unit Task.

Our data suggest that the transparency of a child's count list likely does affect how readily they extract productive counting rules, but that this is not the only factor, and that training amount may overwhelm differences in transparency when grammatical differences in count structure are small. First, we found that when languages exhibited smaller differences in count list structure between them, as in the case of Slovenian, English, and Cantonese, these differences were not the best predictors of performance. For example, whereas English exhibits exceptions in the teens (*eleven, twelve, thirteen*) as well as on decade labels (*twenty, thirty, fifty*), Slovenian only has exceptions in the teens and one decade label (twenty), but nevertheless Slovenian children performed worse than US children on most tasks. One obvious account of why this might be is that Slovenian children likely receive much less exposure to counting in their preschool years, as evidenced by their very low Initial Highest Counts relative to English and Cantonese. Compatible with this, previous studies find that whereas US 5-year-olds typically have an initial highest count up to about 40 or 50 on average, Slovenian children at the same age can only count to about 10 before making their first error (Almoammer et al., 2013; Marušič et al., 2016). Our data also show that Cantonese children don't exhibit a general advantage over US children, particularly when we account for differences in working memory and Initial Highest Count. On the other hand, languages with more extreme morphological complexities were associated with lower performance on these tasks in comparison to more transparent languages, despite similar Initial Highest Count performance. We found that very few learners of Hindi and Gujarati were able to count very far beyond their Initial Highest Count, even when given a prompt, and that these groups performed significantly worse on the Unit Task than English-speaking US children.[8] These data collectively support the idea that counting transparency likely plays a role in children's ability to extract productive counting rules when differences in transparency are significant, and when other factors, like training amount, don't compensate for differences in counting structure. Taken together, our results suggest that when making cross-cultural comparisons, we can't assume that differences between cultures are straightforwardly predicted by differences in grammatical structure, since other factors, like amount of input, surely also play a role.

One important lesson from these studies is that a common measure of children's counting ability - i.e., their Initial Highest Count - often underestimates children's counting ability, and provides a less powerful predictor of other abilities than do measures that are more sensitive to counting productivity. Our data show that, even within a language, two children who have an Initial Highest Count of 30 may have qualitatively different understanding of counting and the rules that govern it. One type of such a child may have rote memorized all numbers up to ~30, without having extracted any rules to describe the count structure. Such children were frequent in Hindi and Gujarati, and surprisingly, in Cantonese, where children were often able to count quite high without error, yet still performed poorly on the Next Number task. Another type of child who counted up to ~30, however, may have noticed the recurrence of the numbers 1-9 in each of

[8] A teacher survey found that teachers expected their Hindi and Gujarati students to be able to count as high as students in English medium schools, and also that Hindi, Gujarati, and English medium school teachers spend similar amounts of time on number and counting instruction in the classroom. Further, Hindi and Gujarati children often recited their count list as a memorized routine, indicating a high level of (rote) training. Despite their relatively high exposure to the count list, however, we still found very few Resilient counters in these two languages, and perhaps because of this, observed much lower mean performance on both the Unit and Next Number tasks.

the first two decades and may have used this observation to extract a rule, stopping at 30 due to a random error - a common pattern in English, Cantonese, and Slovenian. Our methods allowed us to differentiate these two types of child by providing a prompt and asking whether they could continue: Children who have memorized up to 30 should have no idea what to do next, whereas children who have a rule may be able to recover and count up. Not only did we find this to be the case - that a large percentage of children who initially counted to a relatively small number in fact had a productive counting rule (e.g., about 50% of English-speaking children) - we also found that this difference between children who could count-up vs. those who could not (i.e., what we termed "Resilience") was predictive of performance on both the Next Number task and the Unit Task. These data provide strong evidence that, while a child's Initial Highest Count is generally correlated with other measures of productivity, this measure cannot alone be interpreted as evidence of whether they've acquired a productive rule.

Our findings suggest that, when a language provides the basis for acquiring a productive rule for describing a count list's base system, children can learn this rule and use it to generate very large numbers. In the Introduction, we speculated that acquiring such a productive rule might provide the basis for learning not only how to count, but also for discovering the recursive nature of the integers themselves (i.e., the concepts that are denoted by number words). Previous accounts of number word learning hypothesized that children might acquire knowledge of the successor function - and thus of integer concepts - using a form of analogical mapping defined over small number words (Carey, 2004, 2009; Gentner, 2010; Wynn, 1992). On this view, a child who has learned a handful of number words might notice an analogy "between the magnitudinal relationships of their own representations of numerosities, and the positional relationships of the number words" (Wynn, 1992, p.250), such that "she is in the position to make the crucial induction: For any word on the list whose quantificational meaning is known, the next word on the list refers to a set with another individual added" (Carey, 2004, p.67). Originally, this idea was proposed as an account of how children might become CP-knowers, since learning this type of analogical mapping would allow children to use a counting procedure to accurately give sets of any size within their count list. Although we now know that children fail to acquire successor function knowledge at this stage (Cheung et al., 2017; Davidson et al., 2012; Spaepen et al., 2018), it remains possible that this model might still explain how older children learn to interpret their count list. However, a critical problem with this idea is that, for children who have a finite count list, an inductive inference that applies to "all" numbers need not take the form of a recursive function that can generate an infinite number of numbers. Whereas the Peano-Dedekind axioms state that *every* number has a successor, the analogical mapping hypothesis described by Carey and others generates a much weaker inductive inference - i.e., that for all numbers, the successor of N in the count list has a cardinal value of N+1. Whereas a child who has acquired a productive morphological rule for generating indefinitely many number words might take "all" numbers to be unbounded - and thus requiring a recursive rule - a child who knows only 30-40 number words and who believes that numbers are finite — as most 3- and 4-year-olds do (Cheung et al., 2017; Evans, 1983; Gelman, 1980; Hartnett & Gelman, 1998) — might have no basis for inducing a fully recursive successor function (for a related point, see Rips, Asmuth, & Bloomfield, 2006). Said otherwise, if children acquire the successor function by making an analogy between the structure of the count list and the relations between integer concepts, then learning that count words are generated by recursive rules might provide a basis for inferring that the integers, themselves, are governed by recursive rules.

Much remains to be discovered about the process by which children acquire successor function knowledge. For example, while it is plausible that rules governing counting might be analogically extended to cardinal values, more evidence of this is needed. Many alternative hypotheses are possible, including the possibility that explicit arithmetic training - e.g., on problems like 2+1, 3+1, 12+1, etc., forms the basis for an inductive inference supporting the successor function, and happens to co-occur developmentally with greater counting abilities (see Barner, 2017; Secada et al., 1983). Alternatively, performance on the Unit Task may simply be easier once children have acquired a productive counting rule, allowing them to deploy working memory resources previously devoted to tracking their position in the count list to the problem of reasoning about the corresponding set operations. Although all of our cross-linguistic models controlled for working memory, it remains possible that more subtle measures of how working memory is deployed during the Unit Task might find differences in overall load when children use a memorized list vs. one that is governed by rules. Future studies should explore these questions, while also investigated a broader range of languages, using both correlational and experimental designs.

Acknowledgements

We would like to thank Pierina Cheung, Kaiqi Guo, Calvin Lee, Sneha Patel, Shrusti Shah, Soham Shah, Ashlie Pankonin, Sara Lee, Samuel Lucero, Hortensia Flores, Sonora Grimstead, Kaitlyn Seifert, Samuel Beech, and Sarah Valadez for their assistance with data collection. We are grateful to our testing sites in Hong Kong, Nova Gorica, Vadodara, and San Diego, and all of our participants for their time. This work was supported by the National Science Foundation [BCS-1749518, DGE-1321851].

References

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., O'Donnell, T., & Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *Proceedings of the National Academy of Sciences*, *110*(46), 18448-18453.

Barner, D. (2017). Language, procedures, and the non-perceptual origin of number word meanings. *Journal of child language*, *44*(3), 553-590.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.

Barth, H., Starr, A., & Sullivan, J. (2009). Children's mappings of large number words to numerosities. *Cognitive Development*, *24*(3), 248-264.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Berger, H. (1992). Modern Indo-aryan. *Indo-European numerals*, *57*, 243-287.

Bright, W. 1969). *Hindi Numerals*. Washington, D.C.: Distributed by ERIC Clearinghouse, https://eric.ed.gov/?id=ED034175

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59-68.

Carey, S. (2009). Where our number concepts come from. *The Journal of philosophy*, *106*(4), 220.

Cheung, P., Rubenson, M., & Barner, D. (2017). To infinity and beyond: Children generalize the successor function to all possible numbers years after learning to count. *Cognitive psychology*, *92*, 22-36.

Chomsky N. (1965). Aspects of the Theory of Syntax , Vol. 11. Cambridge, MA: MIT Press.

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction?. *Cognition*, *123*(1), 162-173.

Dowker, A., Bala, S., & Lloyd, D. (2008). Linguistic influences on mathematical development: How important is the transparency of the counting system?. *Philosophical Psychology*, *21*(4), 523-538.

Evans, D. W. (1983).Understanding zero and infinity in the early school years. Unpublished doctoral dissertation, University of Pennsylvania.

Fuson, K. C. (1988). *Children's counting and concepts of number.* New York: Springer-Verlag.

Fuson, K. C., & Kwon, Y. (1992). Korean children's understanding of multidigit addition and subtraction. *Child development*, *63*(2), 491-506.

Fuson, K. C., Richards, J., & Briars, D. J. (1982). The acquisition and elaboration of the number word sequence. In *Children's logical and mathematical cognition* (pp. 33-92). Springer, New York, NY.

Geary, D. C. (2018). Growth of symbolic number knowledge accelerates after children understand cardinality. Cognition, 177, 69-78.

Geary, D. C., vanMarle, K., Chu, F. W., Rouder, J., Hoard, M. K., & Nugent, L. (2018). Early Conceptual Understanding of Cardinality Predicts Superior School-Entry Number-System Knowledge. Psychological science, 29(2), 191-205.

Gelman, R., & Gallistel, C. R. (1978). The child's concept of number. *Cambridge, MA: Harvard*.

Gelman, R. (1980). What young children know about numbers. Educational Psychologist, 15, 54-68.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752-775.

Gould, P. (2017). Mapping the acquisition of the number word sequence in the first year of school. *Mathematics Education Research Journal*, *29*(1), 93-112.

Gunderson, E. A., Spaepen, E., & Levine, S. C. (2015). Approximate number word knowledge before the cardinal principle. *Journal of Experimental Child Psychology*, *130*, 35-55.

Hartnett, P., & Gelman, R. (1998). Early understandings of numbers: Paths or barriers to the construction of new understandings?. *Learning and instruction*, *8*(4), 341-374.

Hurford, J. R. (1987). Language and number: The emergence of a cognitive system. Oxford: Basil Blackwell.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395-438.

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*(2), 130-169.

Marchand, E., & Barner, D. (2018). Analogical Mapping in Numerical Development. In *Language and Culture in Mathematical Cognition* (pp. 31-47). Academic Press.

Marušič, F., Plesničar, V., Razboršek, T., Sullivan, J., & Barner, D. (2016). Does grammatical structure accelerate number word learning? Evidence from learners of dual and non-dual dialects of Slovenian. *PloS one*, *11*(8), e0159208.

Miller, K. F., Kelly, M., & Zhou, X. (2005). Learning mathematics in China and the United States: Cross-cultural insights into the nature and course of preschool mathematical development. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 163-177). New York, NY, US: Psychology Press.

Miller, K. F., Smith, C. M., Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science*, *6*(1), 56-60.

Miller, K. F., & Stigler, J. W. (1987). Counting in Chinese: Cultural variation in a basic cognitive skill. *Cognitive Development*, *2*(3), 279-305.

Miura, I. T., Kim, C. C., Chang, C. M., & Okamoto, Y. (1988). Effects of language characteristics on children's cognitive representation of number: Cross-national comparisons. *Child Development*, 1445-1450.

Miura, I. T., & Okamoto, Y. (1989). Comparisons of U.S. and Japanese First Graders' Cognitive Representation of Number and Understanding of Place Value. *Journal of Educational Psychology*, *81*(1), 109–114.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213.

Pan, Y., Gauvain, M., Liu, Z., & Cheng, L. (2006). American and Chinese parental involvement in young children's mathematics learning. *Cognitive Development*, *21*(1), 17-35.

Rips, L. J., Asmuth, J., & Bloomfield, A. (2006). Giving the boot to the bootstrap: How not to learn the natural numbers. Cognition, 101(3), B51-B60.

Rule, J., Dechter, E., & Tenenbaum, J. B. (2015). Representing and Learning a Large System of Number Concepts with Latent Predicate Networks. In *CogSci* (pp. 2051-2056).

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662-674.

Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models, *The R Journal*, 8/1, pp. 205-233. https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf

Secada, W. G., Fuson, K. C., & Hall, J. W. (1983). The transition from counting-all to counting-on in addition. *Journal for Research in Mathematics Education*, 47-57.

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012). Early predictors of high school mathematics achievement. Psychological science, 23(7), 691-697.

Siegler, R. S., & Robinson, M. (1982). The development of numerical understandings. In *Advances in child development and behavior* (Vol. 16, pp. 241-312). JAI.

Spaepen, E., Gunderson, E.A., Gibson, D.G., Goldin-Meadow, S., & Levine, S.C. (2018) Meaning before order: Cardinal principle knowledge predicts improvement in understanding the successor principle and exact ordering. Cognition, 180, 59-81.

Towse, J., & Saxton, M. (1998). Mathematics across national boundaries: Cultural and linguistic perspectives on numerical competence.

von Humboldt, W. F. (1999). On language: On the diversity of human language construction and its influence on the mental development of the human species. Cambridge University Press.

Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive psychology*, *83*, 1-21.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. Educational Researcher, 43(7), 352-360.

Wechsler, D., & Psychological Corporation. (2012). *WPPSI-IV: Wechsler preschool and primary scale of intelligence -- fourth edition*. Bloomington, MN: Pearson, Psychological Corporation.

Wright, R. J. (1994). A study of the numerical development of 5-year-olds and 6-year-olds. *Educational Studies in Mathematics*, *26*(1), 25-44.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155-193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive psychology*, *24*(2), 220-251.

Yang, C. (2016). The linguistic origin of the next number. *Manuscript*

Do Children Use Language Structure to Discover the Recursive Rules of Counting?

Schneider, Sullivan, Marušič, Žaucer, Biswas, Mišmaš, Plesničar, & Barner

Supplementary Online Material

Table of Contents

1. Unit Task: Model building process and interim results

We constructed Unit Task generalized linear mixed effects models using the 'lme4' package in R (Bates et al., 2015). The base model had the following formula: Correct ~ Within/Outside IHC + Item Magnitude + Age + (1|Subject). Continuous variables were scaled and centered to facilitate model fit. We first constructed four individual models with the following candidate measures of productivity: (1) Counting Resilience; (2) Final Highest Count; (3) Initial Highest Count; and (4) Highest Contiguous Next Number. Within each language, we conducted a Likelihood Ratio Test between each of Models 1-4 and the base model to determine whether a candidate productivity measure was significantly related to Unit Task performance.

As preregistered, after determining which productivity measures were individually predictive of successor knowledge, we constructed a single large model within each language using hierarchical model comparison. For each language we built a large model containing the productivity measure associated with the lowest AIC. We then added the other productivity measures which significantly predicted Unit Task performance in that language in order of increasing AIC. We performed a Likelihood Ratio Test with the addition of each new term, and retained that term on the basis of a significant χ^2 statistic.

1.1. Cantonese. Our individual models indicated three productivity measures significantly improved the fit of the base model (Table 1): Final Highest Count ($\chi^2_{(1)} = 11.66$, $p = .0006$); Initial Highest Count ($\chi^2_{(1)} = 38.91$, $p < .0001$); and Highest Contiguous Next Number ($\chi^2_{(1)} = 11.83$, $p = .0006$). Resilience was not significantly related to Unit Task performance ($\chi^2_{(1)} = 0.05$, $p = .82$). The base for our large model contained Initial Highest Count, which yielded the lowest AIC in our individual Likelihood Ratio Tests. Adding Highest Contiguous Next Number did not significantly improve the fit of this model ($\chi^2_{(1)} = 1.86$, $p = .17$), nor did Final Highest Count ($\chi^2_{(1)} = 1.39$, $p = .24$).

Cantonese	***Coefficient estimates (β)***				
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC	Model 4: HCNN
(Intercept)	0.53***	0.51**	0.58***	0.69***	0.56***
Resilient	—	0.05	—	—	—
FHC	—	—	0.44**	—	—
IHC	—	—	—	0.81***	—
HCNN	—	—	—	—	0.39**
Trial Within IHC	0.26	0.26	0.15	-0.05	0.21
Item Magnitude	-0.39***	-0.39***	-0.44***	-0.52***	-0.41***
Age	0.56***	0.55***	0.26*	0	0.37***
AIC	1626.2	1628.1	1616.5	1589.3	1616.3
Conditional R^2	0.261	0.261	0.262	0.266	0.265

Table 1. Base and individual productivity model regression models for predicting Unit Task performance in Cantonese. Coefficients significance was calculated using the standard normal approximation to the *t* distribution (Barr, Levy, Scheepers, & Tily, 2013); **p* < .05; ***p* < .01; ****p* < .001. Conditional R^2 calculated using both fixed and random effects (Nakagawa, Johnson, & Schielzeth, 2017).

1.2. Slovenian. Our individual models indicated all four productivity measures significantly improved the fit of the base model (Table 2): Resilience ($\chi^2_{(1)} = 15.97$, $p < .0001$); Final Highest Count ($\chi^2_{(1)} = 25.13$, $p < .0001$); Initial Highest Count ($\chi^2_{(1)} = 14.59$, $p = .0001$); and Highest Contiguous Next Number ($\chi^2_{(1)} = 26.23$, $p < .0001$). The base for our large model contained Highest Contiguous Next Number, which yielded the lowest AIC in our individual Likelihood

Ratio Tests. The addition of Final Highest Count significantly improved the fit of this model ($\chi^2_{(1)} = 8.03$, $p = .005$). Adding Resilience to a model containing both Final Highest Count and Highest Contiguous Next Number did not improve its fit ($\chi^2_{(1)} = 0.07$, $p = .80$). Finally, Initial Highest Count did not explain unique variance when added to this model ($\chi^2_{(1)} = 0.02$, $p = .89$).

Slovenian	***Coefficient estimates (β)***				
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC	Model 4: HCNN
(Intercept)	0.15	-0.08	0.21*	0.21*	0.19
Resilient	—	0.95***	—	—	—
FHC	—	—	0.57***	—	—
IHC	—	—	—	0.44***	—
HCNN	—	—	—	—	0.53***
Trial Within IHC	0.43*	0.35	0.27	0.27	0.35
Item Magnitude	-0.28***	-0.31***	-0.33***	-0.33***	-0.31***
Age	0.41***	0.23*	0.12	0.25*	0.24**
AIC	1436.7	1422.7	1413.6	1424.1	1412.5
Conditional R^2	0.208	0.212	0.216	0.217	0.218

Table 2. Base and individual productivity model regression models for predicting Unit Task performance in Slovenian. **p* < .05; ***p* < .01; ****p* < .001.

1.3. English (US). Our individual models indicated three productivity measures significantly improved the fit of the base model (Table 3): Final Highest Count ($\chi^2_{(1)} = 18.17$, $p < .0001$); Initial Highest Count ($\chi^2_{(1)} = 48.78$, $p < .0001$); and Highest Contiguous Next Number ($\chi^2_{(1)} = 35.66$, $p < .0001$). The base for our large model contained Initial Highest Count, which yielded the lowest AIC in our individual Likelihood Ratio Tests. Adding Highest Contiguous Next Number significantly improved the fit of this model ($\chi^2_{(1)} = 8.65$, $p = .003$). The addition of Final Highest Count to a model containing both Highest Contiguous Next Number and Initial Highest Count did not improve the fit of this model ($\chi^2_{(1)} = 0.10$, $p = .75$).

English (US)	***Coefficient estimates (β)***				
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC	Model 4: HCNN

(Intercept)	0.43	0.32	0.47***	0.57***	0.50***
Resilient	—	0.28	—	—	—
FHC	—	—	0.64***	—	—
IHC	—	—	—	1.01***	—
HCNN	—	—	—	—	0.81***
Trial Within IHC	0.35	0.34	0.24	0.08	0.27
Item Magnitude	-0.41***	-0.41***	-0.44***	-0.49***	-0.43***
Age	0.69***	0.62***	0.25	0.10	0.26*
AIC	1510.3	1511.2	1494.1	1463.5	1476.7
Conditional R^2	0.333	0.334	0.339	0.360	0.352

Table 3. Base and individual productivity model regression models for predicting Unit Task performance in US English. *$p < .05$; **$p < .01$; ***$p < .001$.

1.4. Hindi. The results of our individual model comparisons indicated that three candidate measures of productivity significantly improved the base model (Table 4): Final Highest Count ($\chi^2_{(1)} = 21.60$, $p < .0001$); Initial Highest Count ($\chi^2_{(1)} = 28.36$, $p < .0001$); and Highest Contiguous Next Number ($\chi^2_{(1)} = 27.63$, $p < .0001$). The base for our large model contained Initial Highest Count, which was associated with the lowest AIC in our individual model comparisons. The addition of Highest Contiguous Next Number to this model significantly improved its fit ($\chi^2_{(1)} = 9.91$, $p = .002$). The addition of Final Highest Count to a model containing both Initial Highest Count and Highest Contiguous Next Number did not significantly improve its fit ($\chi^2_{(1)} = 0.93$, $p = .34$).

Hindi	***Coefficient estimates (β)***				
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC	Model 4: HCNN
(Intercept)	-0.54***	-0.59***	-0.51***	-0.49***	-0.50***
Resilient	—	0.61	—	—	—
FHC	—	—	0.45***	—	—
IHC	—	—	—	0.53***	—
HCNN	—	—	—	—	0.51***
Trial Within IHC	0.14	0.14	0.07	0.01	0.12
Item Magnitude	-0.36***	-0.36***	-0.38***	-0.40***	-0.37***

Age	0.50***	0.48***	0.41***	0.32***	0.37***
AIC	1344.7	1343.5	1325.1	1318.4	1319.1
Conditional R^2	0.211	0.211	0.213	0.215	0.214

Table 4. Base and individual productivity model regression models for predicting Unit Task performance in Hindi. *p < .05; **p < .01; ***p < .001.

1.5. Gujarati. Our individual model comparisons indicated that none of our candidate productivity measures significantly improved the base model (Table 5).

Gujarati	***Coefficient estimates (β)***				
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC	Model 4: HCNN
(Intercept)	-0.35***	-0.38***	-0.34***	-0.34***	-0.34***
Resilient	—	0.20	—	—	—
FHC	—	—	0.09	—	—
IHC	—	—	—	0.11	—
HCNN	—	—	—	—	0.11
Trial Within IHC	0.54**	0.53**	0.50*	0.48*	0.51**
Item Magnitude	-0.16	-0.17	-0.17	-0.18*	-0.17
Age	0.19*	0.18*	0.18*	0.17*	0.18*
AIC	1272.7	1274.0	1273.7	1273.0	1273.1
Conditional R^2	0.093	0.094	0.093	0.093	0.093

Table 5. Base and individual productivity model regression models for predicting Unit Task performance in Gujarati. *p < .05; **p < .01; ***p < .001.

1.6. Cross-linguistic models. The results of our individual model comparisons revealed that only Highest Contiguous Next Number significantly improved the base model in both Experiment 1 (Table 6, $\chi^2_{(1)} = 22.70$, $p < .0001$) and Experiment 2 (Table 7, $\chi^2_{(1)} = 13.18$, p =.0003).

	Comparison to Cantonese				**Comparison to Slovenian**			
	Coefficient Estimates (β)				***Coefficient Estimates (β)***			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: HCNN	Base	Model 1: Resilience	Model 2: FHC	Model 3: HCNN
(Intercept)	0.21*	0.14	0.19	0.31**	0.57***	0.51***	0.57***	0.50***
Resilient	—	0.15	—	—	—	0.15	—	—

FHC	—	—	0.15	—	—	—	0.15	—
HCNN	—	—	—	0.34***	—	—	—	0.34***
Cantonese	—	—	—	—	-0.35*	-0.36*	-0.38*	-0.19
Slovenian	0.35*	0.36*	0.38*	0.20	—	—	—	—
English (US)	0.60***	0.61***	0.61***	0.43**	0.25	0.24	0.24	0.24
IHC	0.68***	0.67***	0.58***	0.54***	0.77***	0.72***	0.62**	0.55**
Trial Within IHC	0.10	0.10	0.10	0.10	0.10	0.10	010	0.10
Item Magnitude	-0.44***	-0.44***	-0.44***	-0.44***	-0.44***	-0.44***	-0.44***	-0.44***
Age	0.16	0.13	0.12	0.10	0.16	0.13	0.12	0.10
WPPSI	0.09	0.08	0.08	0.07	0.09	0.08	0.08	0.07
Cantonese: IHC	—	—	—	—	-0.10	-0.06	-0.04	-0.01
Slovenian: IHC	0.10	0.06	0.04	0.01	—	—	—	—
English (US): IHC	0.31*	0.30*	0.30*	0.20	0.21	0.24	0.25	0.19
AIC	4400.02	4400.5	4400.00	4379.3				
Conditional R^2	0.292	0.292	0.291	0.293				

Table 6. Base and individual productivity model regression models for predicting Unit Task performance in cross-linguistic analyses with Cantonese, Slovenian, and US English, with Cantonese (left) and Slovenian (right) selected as reference groups. **p* < .05; ***p* < .01; ****p* < .001.

	Coefficient Estimates (β)			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: HCNN
(Intercept)	0.45***	0.44***	0.45***	0.43***
Resilient	-	0.02	-	-
FHC	-	-	-0.01	-
HCNN	-	-	-	0.26***
Hindi	-0.71***	-0.71***	-0.71***	-0.70***
Gujarati	-0.75***	-0.75***	-0.75***	-0.72***
IHC	0.59***	0.59***	0.60***	0.45***
Trial Within IHC	0.15	0.15	0.15	0.16
Item Magnitude	-0.37***	-0.37***	-0.37***	-0.37***
Age	0.30**	0.30**	0.30**	0.27**
WPPSI	0.12*	0.12*	0.12*	0.10
Hindi: IHC	0.56**	0.56**	0.56**	0.45*
Gujarati: IHC	-0.37*	-0.38*	-0.37*	-0.42*
AIC	3951.63	3953.6	3953.6	3940.5
Conditional R^2	0.281	0.281	0.281	0.284

Table 7. Base and individual productivity model regression models for predicting Unit Task performance in cross-linguistic analyses with Hindi, Gujarati, and US English, with US English as the reference group. *$p < .05$; **$p < .01$; ***$p < .001$.

2. Next Number Task: Model building process and interim results

We constructed our Next Number Task models with the same process we used in our Unit Task analyses. The base model had the following formula: Correct ~ Within/Outside IHC + Item Magnitude + Age + (1|Subject). Continuous variables were scaled and centered to facilitate model fit. We first built three individual models with the following candidate measures of productivity: (1) Counting Resilience; (2) Final Highest Count; and (3) Initial Highest Count.

2.1. Cantonese. Individual models indicated two significant predictors of Next Number performance (Table 8): Final Highest Count ($\chi^2_{(1)} = 34.33$, $p < .0001$) and Initial Highest Count ($\chi^2_{(1)} = 68.63$, $p < .0001$). The base for our large model predicting Next Number performance contained Initial Highest Count, which was associated with the lowest AIC in our individual model comparisons. The addition of Final Highest Count to this model did not significantly improve its fit ($\chi^2_{(1)} = 0.37$, $p = .54$).

Cantonese	***Coefficient estimates (β)***			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC
(Intercept)	-0.69***	-0.84**	-0.62***	-0.47**
Resilient	—	0.31	—	—
FHC	—	—	1.20***	—
IHC	—	—	—	1.62***
Trial Within IHC	1.20***	1.19***	1.03***	0.75**
Item Magnitude	-0.92***	-0.92***	-0.99***	-1.12***
Age	1.49***	1.44***	0.67***	0.34*
AIC	1203.5	1204.7	1171.2	1136.9
Conditional R^2	0.674	0.674	0.678	0.677

Table 8. Base and individual productivity model regression models for predicting Next Number Task performance in Cantonese. *$p < .05$; **$p < .01$; ***$p < .001$.

2.2. Slovenian. Individual models revealed that all three candidate measures were significantly related to Next Number performance (Table 9): Resilience ($\chi^2_{(1)} = 40.26, p < .0001$); Final Highest Count ($\chi^2_{(1)} = 72.93, p < .0001$); and Initial Highest Count ($\chi^2_{(1)} = 36.19, p < .0001$). We constructed the base for our large model with Final Highest Count, which was associated with the lowest AIC in our individual model comparisons. The addition of neither Resilience ($\chi^2_{(1)} = 0.004, p = .99$) nor Initial Highest Count ($\chi^2_{(1)} = 0.10, p = .76$) significantly improved the fit of this model.

Slovenian	***Coefficient estimates (β)***			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC
(Intercept)	-0.63**	-1.35***	-0.31*	-0.44*
Resilient	—	2.86***	—	—
FHC	—	—	1.83***	—
IHC	—	—	—	1.37***
Trial Within IHC	1.46***	1.37***	1.19***	1.26***
Item Magnitude	-0.95***	-1.01***	-1.09***	-1.04***
Age	1.54***	0.89***	0.47**	0.97***
AIC	901.56	863.31	830.63	867.38
Conditional R^2	0.698	0.693	0.698	0.706

Table 9. Base and individual productivity model regression models for predicting Next Number Task performance in Slovenian. **p* < .05; ***p* < .01; ****p* < .001.

2.3. English (US). Our individual models indicated that all three candidate measures significantly predicted Next Number performance (Table 10): Resilience ($\chi^2_{(1)} = 14.76, p = .0001$); Final Highest Count ($\chi^2_{(1)} = 44.00, p < .0001$); and Initial Highest Count ($\chi^2_{(1)} = 30.52, p < .0001$). We constructed our large model with Final Highest Count, which was associated with the lowest AIC in our individual model comparisons. in the base. The addition of Initial Highest Count to this base only marginally improved the fit of the model ($\chi^2_{(1)} = 3.73, p = .053$), and Resilience did not explain any additional variance ($\chi^2_{(1)} = 0.70, p = .40$).

English (US)	***Coefficient estimates (β)***			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC
(Intercept)	-0.50*	-1.13***	-0.42*	-0.32
Resilient	—	1.59***	—	—
FHC	—	—	1.49***	—
IHC	—	—	—	1.33***
Trial Within IHC	1.81***	1.81***	1.64***	1.57***
Item Magnitude	-0.59***	-0.59***	-0.66***	-0.69***
Age	1.76***	1.33***	0.69**	0.99***
AIC	1054.78	1041.8	1012.6	1026.1
Conditional R^2	0.716	0.713	0.718	0.737

Table 10. Base and individual productivity model regression models for predicting Next Number Task performance in US English. *$p < .05$; **$p < .01$; ***$p < .001$.

2.4. Hindi. Individual model comparisons revealed that all three candidate productivity measures significantly improved the fit of the base model (Table 11): Resilience ($\chi^2_{(1)} = 4.55$, $p = .03$); Final Highest Count ($\chi^2_{(1)} = 27.38$, $p < .0001$); and Initial Highest Count ($\chi^2_{(1)} = 33.41$, $p < .0001$). The base for the large model included Initial Highest Count, which resulted in the lowest AIC in our individual model comparisons. The addition of Final Highest Count only marginally improved the fit of this model ($\chi^2_{(1)} = 2.77$, $p = .10$), while the addition of Resilience did not produce a better fit to the data ($\chi^2_{(1)} = 0.37$, $p = .54$).

Hindi	***Coefficient estimates (β)***			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC
(Intercept)	-1.97***	-2.12***	-1.91***	-1.88***
Resilient	—	1.79*	—	—
FHC	—	—	1.25***	—
IHC	—	—	—	1.41***
Trial Within IHC	1.85***	1.85***	1.76***	1.68***
Item Magnitude	-0.85***	-0.86***	-0.89***	-0.92***
Age	0.98***	0.93***	0.76**	0.53*

AIC	853.0	850.45	827.62	821.59
Conditional R^2	0.701	0.701	0.713	0.714

Table 11. Base and individual productivity model regression models for predicting Next Number Task performance in Hindi. *$p < .05$; **$p < .01$; ***$p < .001$.

2.5. Gujarati. Our individual model comparisons indicated that all three candidate measures of productivity significantly predicted Next Number performance (Table 12): Resilience ($\chi^2_{(1)} = 10.06$, $p = .002$), Final Highest Count ($\chi^2_{(1)} = 36.25$, $p < .0001$), and Initial Highest Count ($\chi^2_{(1)} =33.45$, $p < .0001$) all significantly improved the fit of the model in comparison to the base. The large model was constructed using Final Highest Count, which was associated with the lowest AIC in our individual model comparisons, as the base. The addition of Initial Highest Count to this model did not produce a better fit ($\chi^2_{(1)} = 0.87$, $p = .35$). The addition of Resilience, however, did significantly explain additional variance ($\chi^2_{(1)} = 4.12$, $p = .04$).

Gujarati	***Coefficient estimates (β)***			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC
(Intercept)	-1.54***	-1.80***	-1.47***	-1.46***
Resilient	—	2.09***	—	—
FHC	—	—	1.30***	—
IHC	—	—	—	1.25***
Trial Within IHC	2.23***	2.19***	2.03***	1.98***
Item Magnitude	-1.16***	-1.17***	-1.24***	-1.26***
Age	0.48*	0.44	0.37	0.32
AIC	795.5	787.41	761.23	764.02
Conditional R^2	0.695	0.696	0.710	0.710

Table 12. Base and individual productivity model regression models for predicting Next Number Task performance in Gujarati. *$p < .05$; **$p < .01$; ***$p < .001$.

2.6. English (India). Our individual model comparisons indicated that Final Highest Count ($\chi^2_{(1)} = 21.1$, $p < .0001$) and Initial Highest Count ($\chi^2_{(1)} = 19.49$, $p = .0001$) significantly predicted Next Number performance (Table 13). We built our large model with Final Highest Count included in the base, as this measure was associated with the lowest AIC in our individual model comparisons. The addition of Initial Highest Count significantly improved the fit of this model ($\chi^2_{(1)} = 5.88$, $p = .02$).

English (India)	***Coefficient estimates (β)***			
Parameters	Base	Model 1: Resilience	Model 2: FHC	Model 3: IHC
(Intercept)	-0.87***	-1.10***	-0.80***	-0.77***
Resilient	—	0.56	—	—
FHC	—	—	1.27***	—
IHC	—	—	—	1.06***
Trial Within IHC	1.40***	1.41***	1.23***	1.13***
Item Magnitude	-1.25***	-1.25***	-1.31***	-1.37***
Age	1.21***	1.06***	0.25	0.57*
AIC	697.8	698.32	678.68	680.30
Conditional R^2	0.676	0.676	0.678	0.685

Table 13. Base and individual productivity model regression models for predicting Next Number Task performance in Indian English. *$p < .05$; **$p < .01$; ***$p < .001$.

2.7. Cross-linguistic models. In Experiment 1 (Table 14, Cantonese, Slovenian, and US English), our individual model comparisons indicated that both Resilience ($\chi^2_{(1)} = 19.4$, $p < .0001$) and Final Highest Count ($\chi^2_{(1)} = 38.96$, $p < .0001$) significantly improved the fit of the base model. The addition of Resilience to a model containing Final Highest Count did not explain additional variance ($\chi^2_{(1)} = 0.27$, $p = .60$).

	Comparison to Cantonese			**Comparison to Slovenian**		
	Coefficient Estimates (*β*)			**Coefficient Estimates (*β*)**		
Predictors	Base	Model 1: Resilience	Model 2: FHC	Base	Model 1: Resilience	Model 2: FHC
(Intercept)	-1.38***	-1.80***	-1.51***	0.21	-0.14	0.27
Resilient	-	0.88***	-	-	0.88***	-
FHC	-	-	1.05***	-	-	1.05***
Cantonese	-	-	-	-1.58***	-1.66***	-1.77***
Slovenian	1.58***	1.66***	1.77***	-	-	-
English (US)	1.77***	1.78***	1.82***	0.18	0.12	0.04

IHC	1.27***	1.26***	0.65**	1.93***	1.65***	0.88**
Trial Within IHC	1.20***	1.19***	1.18***	1.20***	1.19***	1.18***
Item Magnitude	-0.91***	-0.92***	-0.92***	-0.91***	-0.92***	-0.92***
Age	0.92***	0.75***	0.61***	0.92***	0.75***	0.61***
WPPSI	0.19*	0.17	0.15	0.19*	0.17	0.15
Cantonese: IHC	-	-	-	-0.65*	-0.39	-0.23
Slovenian: IHC	0.65*	0.39	0.23	-	-	-
English (US): IHC	0.26	0.17	0.13	-0.40	-0.21	-0.11
AIC	2986.06	2968.2	2949.1			
Conditional R^2	0.708	0.706	0.702			

Table 14. Base and individual productivity model regression models for predicting Next Number performance in cross-linguistic analyses with Cantonese, Slovenian, and US English, with Cantonese (left) and Slovenian (right) selected as reference groups. **p* < .05; ***p* < .01; ****p* < .001.

In Experiment 2 (Table 15, Hindi, Gujarati, and US English), we again found that both Resilience ($\chi^2_{(1)} = 9.91$, $p = .002$) and Final Highest Count ($\chi^2_{(1)} = 22.68$, $p < .0001$) significantly improved the fit of the base model. The addition of Resilience to a model containing Final Highest Count did not significantly improve its fit ($\chi^2_{(1)} = .64$, $p = .43$). In a *post hoc* analysis with Indian English substituted for US English, we also found that only Final Highest Count significantly improved the fit of the base model ($\chi^2_{(1)} = 8.36$, $p = .004$).

	Comparison to English (US)			**Comparison to English (India)**		
	Coefficient Estimates (*β*)			**Coefficient Estimates (*β*)**		
Predictors	Base	Model 1: Resilience	Model 2: FHC	Base	Model 1: Resilience	Model 2: FHC
(Intercept)	-0.33	-0.74**	-0.63**	-0.97***	-1.19***	-1.33***
Resilient	-	0.98**	-	-	0.50	-
FHC	-	-	0.98***	-	-	0.66**
Hindi	-1.00**	-0.67	-0.53	-0.17	-0.01	0.30
Gujarati	-0.93**	-0.66	-0.53	-0.14	0	0.29
IHC	1.11***	1.01***	0.42*	0.63***	0.62**	0.25
Trial Within IHC	1.73***	1.73***	1.73***	1.66***	1.67***	1.66***
Item Magnitude	-0.90***	-0.90***	-0.91***	-1.21***	-1.21***	-1.21***
Age	0.97***	0.85***	0.80***	0.71**	0.64**	0.57**
WPPSI	0.20	0.20	0.15	0.20	0.18	0.13
Hindi: IHC	1.68***	1.61**	1.48**	2.19***	2.13***	2.05***

Gujarati: IHC	0.75	0.50	0.45	1.22**	1.07**	1.00**
AIC	2558.2	2550.3	2544.2	2146.6	2146.5	2140.2
Conditional R^2	0.738	0.736	0.733	0.717	0.716	0.716

Table 15. Base and individual productivity model regression models for predicting Next Number performance in cross-linguistic analyses with Hindi, Gujarati, US English (left) and Indian English (right). *p < .05; **p < .01; ***p < .001.

3. Analyses of items above and below 100

For all languages we tested, 100 marks the introduction of a new decade label. Children must also learn that the introduction of this new decade label must be appended to the beginning of the count list, and does not replace the other decade labels that came prior. Further, most counting instruction likely does not exceed 100, such that children often have to discover these rules on their own. For many children, these syntactic and morphological changes proved to be quite challenging; in many languages children were able to use prompts to count up to 100, but not beyond. Although we observed a decrease in performance on both the Unit and Next Number tasks with increasing item magnitude in every language (Figure 1), we wished to test whether items above 100 as a group were particularly difficult for children in these tasks. In particular, we explored whether an item being less or greater than 100 accounted for unique variance in children's performance beyond their Initial Highest Count and productive counting knowledge.

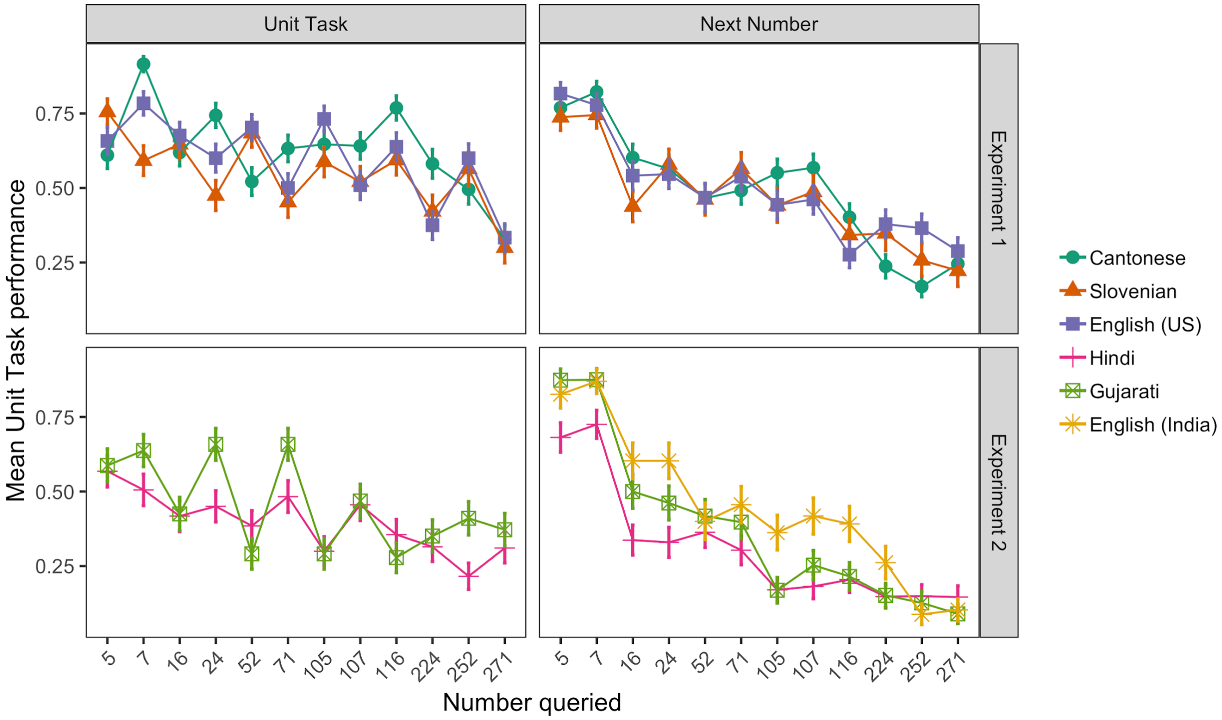

Figure 1. Mean performance by item for the Unit and Next Number tasks in Experiments 1 and 2. Error bars represent standard error of the mean.

To do this, we constructed a base model for both the Unit and Next Number tasks in each language which included the strongest productivity predictors for that task. Because item magnitude and whether the item is above or below 100 overlap, we removed the item magnitude term from our base model. Thus, for each language the generalized linear mixed effects models was: `Correct ~ [Productivity predictors] + Within IHC + Age + (1|Subject)`. We then added a term indicating whether the queried item was above or below 100, and conducted a Likelihood Ratio Test to determine whether this term significantly improved the fit of the model.

3.1. Unit Task. The addition of a term indicating whether a queried item was above or below 100 did not significantly improve the fit of the base model in Cantonese ($\chi^2_{(1)} = .0008$, $p = .98$), or in Slovenian ($\chi^2_{(1)} = 2.94$, $p = .09$). In these two languages, children's performance was more significantly predicted by whether the item was within or outside their Initial Highest Count.

However, adding this term did significantly improve the fit of the model in US English ($\chi^2_{(1)}$ = 6.50, p = .01), Hindi ($\chi^2_{(1)}$ = 13.35, p = .0003), and Gujarati ($\chi^2_{(1)}$ = 11.89, p = .0006), such that performance for items above 100 was significantly worse in each language (US English: β = -0.40, p = .008; Hindi: β = -0.62, p = .0003; and Gujarati: β = -0.59, p = .0006).

3.2. Next Number. As in our Unit Task analysis, the addition of a term indicating whether an item was above or below 100 did not significantly improve the fit of the base model in Cantonese ($\chi^2_{(1)}$ = .12, p = .73); once again, children's performance was better predicted by whether the item was outside their Initial Highest Count than whether the item was above 100. On the other hand, adding this term did significantly improve the fit of the base model in Slovenian ($\chi^2_{(1)}$ = 45.76, p < .0001), US English ($\chi^2_{(1)}$ = 36.91, p < .0001), Hindi ($\chi^2_{(1)}$ = 42.23 p <.0001), and Gujarati ($\chi^2_{(1)}$ = 60.58, p < .0001), such that performance for items above 100 was significantly worse in each language (Slovenian: β =*-1.54, p* < .0001; US English: β = -1.31, p < .0001; Hindi: β = -1.72, p < .0001; and Gujarati: β = -1.94, p < .0001).

Overall, we found some evidence that items above 100 were more difficult for children, particularly in the Next Number task, where children must generate the next number without alternatives. Interestingly, we did not find that items above 100 were more difficult for Cantonese-speaking children, but rather that their performance on both the Unit and Next Number tasks was more closely related to whether an item was within their Initial Highest Count (Unit Task: β = .76, p < .0001; Next Number: β = 2.23, p < .0001). This finding likely reflects that a child's Initial Highest Count was the strongest measure of productivity in Cantonese for both these tasks.

4. Mean Unit and Next Number Task performance by Resilience classification

We tested whether our binary Resilience classification broadly captured differences in children's numerical knowledge by testing mean performance on both the Unit and Next Number tasks within each language. Although we found that graded measures of productive counting knowledge were more strongly predictive of performance on both these tasks, our binary classification nevertheless was individually predictive in most languages using independent samples *t*-tests.

4.1. Unit Task. Resilient counters had significantly higher mean Unit Task performance than Non-Resilient counters in: Slovenian (*t*(97) = 5.68, *p* < .0001); US English (*t*(109) = 3.87, *p* = .0002). The difference in mean performance was marginal in Cantonese (*t*(116) = 1.87, *p* = .06) and Hindi (*t*(89) = 1.99, *p* = .05). There was no difference in mean Unit Task performance between Resilient and Non-Resilient counters in Gujarati (*t*(78) = 1.16, *p* = .25).

4.2. Next Number Task. Supporting the hypothesis that Resilient counters have some mastery of the productive rules underlying number word generation, we found that Resilient counters had significantly higher mean Next Number performance than Non-Resilient counters in all languages: Cantonese (*t*(116) = 3.22, *p* = .002); Slovenian (*t*(97) = 9.21, *p* < .0001); US English (*t*(107) = 7.71, *p* < .0001); Hindi (*t*(89) = 2.27, *p* = .03); Gujarati (*t*(78) = 3.37, *p* = .001); and Indian English (*t*(68) = 3.68, *p* = .0005).

5. Highest Contiguous Next Number

We used Highest Contiguous Next Number, the highest number for which a child could successfully generate the next number in response to a prompt provided that all the previous numbers were correct, as a measure of their knowledge of productive counting rules (Table 16). The lowest Highest Contiguous Next Number possible was 0, meaning that a child made an error on the training trial with 1 (*n* Cantonese = 8; *n* Slovenian = 9; *n* US English = 8; *n* Hindi = 4; *n* Indian English = 5). The highest number possible was 271; this number would indicate perfect performance on this task (*n* Cantonese = 2; *n* Slovenian = 5; *n* US English = 14; *n* Hindi = 6; *n* Gujarati = 2)

	Cantonese	**Slovenian**	**English (US)**	**Hindi**	**Gujarati**	**English (India)**
Overall						
Mean (*SD*)	55 (74.39)	50 (82.56)	66 (98.41)	30 (70.86)	34 (61.55)	43 (65.89)
Median	16	7	7	7	7	12
Resilient						
Mean (*SD*)	74 (85.31)	121 (98.20)	118 (112.61)	83 (113.59)	90 (105.40)	65 (83.72)
Median	24	107	62	30	52	24
Non-Resilient						
Mean (*SD*)	35 (54.42)	25 (59.28)	33 (71.52)	25 (64.16)	25 (46.63)	28 (46.00)
Median	7	7	7	7	7	7

Table 16. Highest Contiguous Next Number by language and Resilience. Means and medians are rounded.

Children identified as Resilient counters had significantly higher Highest Contiguous Next Numbers in all languages in comparison to Non-Resilient counters, indicating that these children have acquired a productive rule for generating number words (Cantonese (*t*(116) = 2.94, *p* = .004); Slovenian (*t*(97) = 5.84, *p* < .0001); US English (*t*(107) = 4.85, *p* < .0001); Hindi (*t*(89) = 2.23, *p* = .03); Gujarati (*t*(78) = 3.49, *p* = .0008); and Indian English (*t*(68) = 2.41, *p* = .02)).

Next, we tested whether children learning a more transparent count list may be able to generate higher Next Numbers even without demonstrating evidence of having acquired productive counting rules. We tested this in both Experiment 1 (Cantonese, Slovenian, and English) and Experiment 2 (Hindi, Gujarati, and US English) by building a linear regression predicting Highest Contiguous Next Number in Non-Resilient counters. As in our other cross-linguistic analyses, we attempted to control for between-group differences by including both an age and working memory term. The formula for this model was: `Highest Contiguous Next Number ~ Language + Age + Working Memory score.`

Reflecting the results of our other cross-linguistic analyses with Cantonese, Slovenian, and US English, we did not find a difference in Highest Contiguous Next Number for Non-Resilient counters between Cantonese and Slovenian (*β* = -14.17, *p* = .17) or between Cantonese and English (*β* = 12.79, *p* = .23). Non-Resilient English-speaking children, on the other hand, had significantly higher Highest Contiguous Next Numbers in comparison to Slovenian children (β = 26.96, p = .01). In Experiment 2 we found that, in comparison to US English, Non-Resilient counters had significantly lower Highest Contiguous Next Numbers in Gujarati (β = -34.64, p = .008) and Hindi (β = -32.32, p = .01). We replicated these results using our Indian English dataset as a reference group (Gujarati: β = -29.83, p = .01; Hindi: β = -27.52, p = .02). Thus, while we found that training, rather than transparency, predicted differences in Non-Resilient counters' performance on this task in three relatively transparent languages (Cantonese, Slovenian, and English), we did find that children learning a more opaque count list (Hindi, Gujarati) were less likely to be able to correctly generate the next number in a sequence without an understanding of their count list's generative syntax.

6. Initial Highest Count

A child's Initial Highest Count is reflective of both the regularity of a count list, as well as frequency of counting training within that language. Thus, we should predict that languages with more regular count lists, and higher levels of count routine exposure (such as Cantonese) should be associated with a greater frequency of high Initial Highest Counts. To test this prediction, we used Gaussian Mixture Modeling to identify and quantify Initial Highest Count distributions in each language (Figure 2). These models were fit using the 'mclust5' package in R (Scrucca, Fop, Murphy, & Raftery, 2017), and aggregated over both Resilient and Non-Resilient Highest Counts. Gaussian Mixture Models (GMM) identify clusters within data (assuming an underlying Gaussian distribution), calculate their mean and variance, and estimate the likelihood of a given point being contained within a given cluster. The number of optimal clusters is selected on the basis of Bayesian Information Criterion. The fits of individual clusters are shown in Table 17.

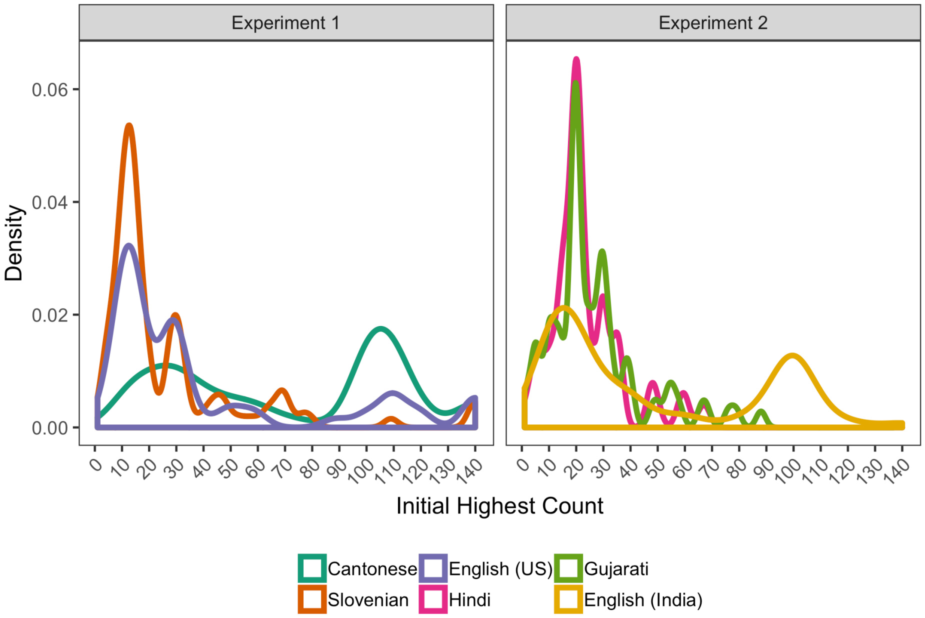

Figure 2. Distribution of Initial Highest Counts in Experiments 1 and 2.

The results of the GMM support our hypothesis that greater levels of training and count list transparency should result in a greater frequency of higher Initial Highest Counts. Cantonese counters' most frequent Initial Highest Count was 106 (Cluster 4), with a number of much lower probability clusters. In languages with either lower levels of exposure (Slovenian) or transparency (English) the most probable clusters were in the teens (Cluster 1), indicating the challenge this decade poses in extracting recursive counting rules. Indian English clusters were similar to US English clusters in both means and probabilities. In both Hindi and Gujarati, cluster membership largely overlapped with our Resilient/Non-Resilient classification.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Cantonese (*n*)	*n* = 20	*n* = 24	*n* =15	*n* = 50	*n* = 20
Mean (*SD*)	17.83 (6.57)	34.18 (6.57)	59.23 (6.57)	106.12 (6.57)	139.89 (6.57)
Probability	0.192	0.189	0.119	0.423	0.077
Proportion Resilient	.20	.46	.73	.52	1.0
Slovenian (*n*)	*n* = 48	*n* = 15	*n* = 16	*n* = 20	
Mean (*SD*)	11.64 (4.52)	13.58 (1.00)	29.73 (1.73)	66.46 (35.05)	
Probability	0.49	0.13	0.15	0.23	—
	.08	0	.44	.75	

Proportion Resilient					
English (US) (*n*)	*n* = 54	*n* = 22	*n* = 35	—	—
Mean (*SD*)	12.55 (3.75)	28.91 (0.40)	89.49 (39.24)		
Probability	0.47	0.19	0.33		
Proportion Resilient	.13	.50	.69		
English (India) (*n*)	*n* = 32	*n* = 15	*n* = 23	—	—
Mean (*SD*)	14.45 (4.94)	33.42 (15.16)	101.01 (11.60)		
Probability	0.40	0.27	0.33		
Proportion Resilient	.22	.67	.48		
Gujarati (*n*)	*n* = 69	*n* = 11	—	—	—
Mean (*SD*)	20.48 (9.22)	56.50 (17.34)			
Probability	0.82	0.18			
Proportion Resilient	.06	.64			
Hindi (*n*)	*n* = 82	*n* = 9	—	—	—
Mean (*SD*)	20.11 (8.40)	55.81 (8.20)			
Probability	0.90	0.10			
Proportion Resilient	.05	.44			

Table 17. Clusters for IHC in each language. Each cluster identifies a mode within the IHC distribution for each language. Mean and SD refer to the mean and SD of that cluster, while probability indicates the likelihood of a given point within that dataset falling into that cluster. Proportion Resilient indicates the proportion of counters in that cluster who met the criteria for Resilience.

Next, we tested whether children need to count higher in languages with less transparent count lists in order to acquire a recursive rule (Yang, 2016) by constructing a linear regression predicting Initial Highest Counts for Non-Resilient counters by language, controlling for age and working memory. Contra this prediction, in Experiment 1 we found that Cantonese Non-Resilient counters had significantly higher IHCs than Slovenian (β = 46.26, p < .0001) and English Non-Resilient counters (β = 20.51, p < .0001). We found again that Non-Resilient counters in more transparent languages were able to count significantly higher before acquiring a productive rule in Experiment 2. US English Non-Resilient counters' had significantly higher Initial Highest Counts in comparison to both Hindi (β = 19.04, p < .0001) and Gujarati (β = 19.40, p < .0001). We replicated this finding in a within-culture comparison: Indian English Non-Resilient counters had significantly higher Initial Highest Counts than both Hindi (β = 27.79, p < .0001) and Gujarati (β = 27.57, p < .0001) Non-Resilient counters. While the significantly lower Initial Highest Counts of Hindi and Gujarati indicate that count list morphology plays a significant role in extractive recursive counting rules, these findings suggests that this is not purely the result of grammatical structure, as Cantonese-speaking children were able to count quite high prior to demonstrating productive counting knowledge.

7. Counting errors

Previous work (Fuson, 1988; Miller & Stigler, 1987; Siegler & Robinson, 1982) has found that children's errors in the count routine are nonrandom, and that many children are more likely to make errors on decade transitions, which require remembering a new decade label. This pattern of errors would suggest that these children have mastered the decade structure, but have not yet committed the decade labels to memory. We explored whether Resilient Counters differ from Non-Resilient counters in the pattern of their counting errors, hypothesizing that children who are productive are more likely to make errors which involve the recall of a new decade label rather than errors mid-decade, which would indicate that they have not yet mastered the base-10 system.

We found that Resilient Counters made more frequent use of more prompts than Non-Resilient Counters, but the kinds of errors made by both counters varied across languages (Figure 3). In Cantonese and English, a higher proportion of Resilient counters' errors (about 70%) were at decade transitions (e.g., 49 to 50) than either decade-beginning (Cantonese: 12%; English; 13%) or mid-decade (Cantonese: 17%; English: 17%). Resilient counters in Slovenian, however, had their highest proportion of errors for mid-decade numbers (50%), with comparatively fewer errors at decade transitions (35%) or beginnings (15%). Hindi and Gujarati Resilient counters also had more frequent decade-transition errors. In both Hindi and Gujarati, the syntax for a decade transition already incorporates the next decade label. For example, in Hindi the progression from 38-40 would be *adh**tis*** (38)*, untaa**lis*** (39)*, cha**lis*** (40)*,* where *tis* and *lis* mean *thirty* and *forty* respectively. In line with our hypothesis that Resilient Counters are more likely to require prompts when transitioning to a new decade label, we find that the majority (43%) of Gujarati Resilient counters' prompts were provided at these points, while these errors comprised 30% of Hindi Resilient counters' stopping points.

Like Slovenian Resilient counters, however, Hindi and Gujarati Resilient counters were also likely to make errors both in the middle and start of decades. In fact, Resilient Hindi Counters made the highest proportion of errors (45%) mid-decade. This high proportion of mid-decade errors is unsurprising, however, given the irregularity within individual decades: For example, in both Hindi and Gujarati, the decade label for *fifty* alternates between *van* and *pan*. Similarly, Slovenian Resilient counters' relatively higher frequency of mid-decade errors may reflect their lower levels of counting training.

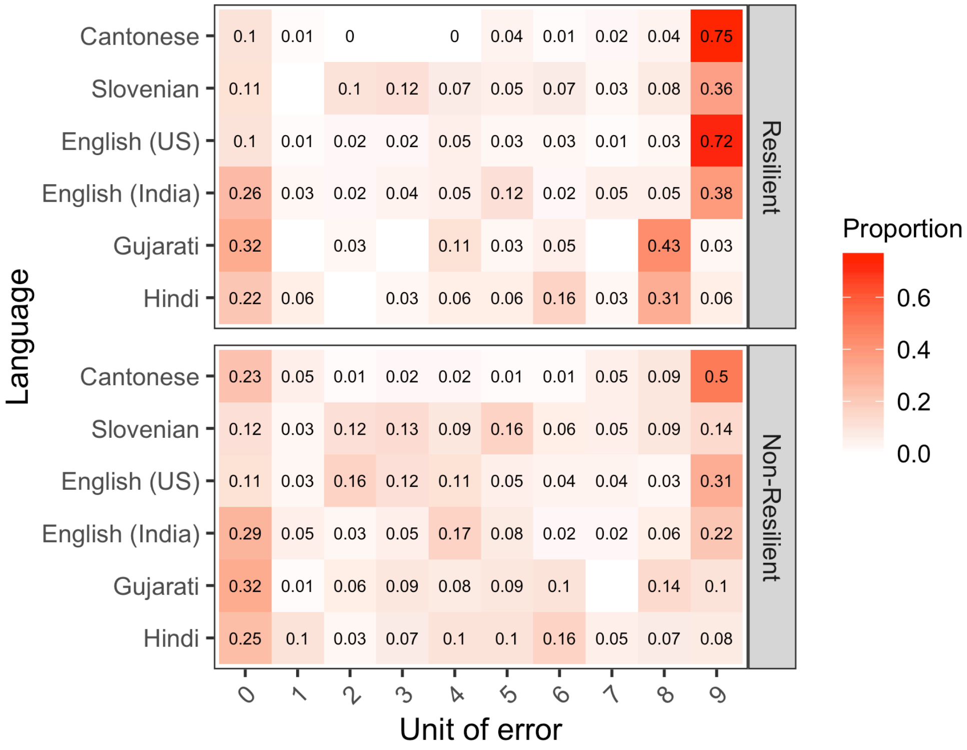

Figure 3. Frequency of unit in which children made an error in each language, grouped by Resilience. Color corresponds to the proportion of total errors made by Resilient or Non-Resilient counters within each language. The unit of error corresponds to a child's last successful count prior to making an error.