Dear Dr. McNeil,

Thanks to you and the three reviewers for your helpful comments on this manuscript. We have made adjustments to the paper to respond to each of these comments, which we detail below.

In particular, we have now done more to describe how we chose our measures (and why), and how we converged upon the cutoffs we used. In brief, we used the infinity interview because it allowed us to ask questions that were general in nature, and because it has been shown to correlate well with other important measures in past studies. Cutoffs (e.g., for defining productivity) were adopted from other studies in our lab that used the same criteria. We note this, and also explain why we believe these criteria are sensible, independent of their use in previous studies.

We have also incorporated your comments. We believe that pattern recognition of some form is almost certainly involved in learning morphological paradigms like the count list, and that this is perhaps a difference of terminology across fields, more than a difference in the substance of our claim. Learning a morphological paradigm involves detecting regularities amidst exceptions, much like domain-general pattern recognition. We have now made this connection in the paper.

We believe that these modifications make the logic of the paper stronger and thank you again for your comments.

Regards,

Junyi Chu

cc:

Rose Schneider

Pierina Cheung

Jessica Sullivan

David Barner

# Overall list of changes

|  |  |  |
| --- | --- | --- |
| **Location** | **Change** | **Suggestion from:** |
| **Introduction** | | |
| p.7 | Added citations (Kaminsky, 2015; Sella, 2020) to the discussion that full successor function understanding comes after CP transition | Editor |
| p.8 | Clarified definition of successor function | Reviewers 1 & 2 |
| p.10 | Added Table 1 to clarify what it means to learn the Decade+Unit Rule, and explained the logic of our hypothesis that noticing recursive syntax can lead children to infer numbers do not end | Reviewers 1 & 2 |
| p.11 | Added text explaining that children may acquire a successor function for generating numbers without appreciating the logical entailment of infinity | Reviewer 2 |
| p.12 | Replaced “understanding how this recursive rule relates to cardinality” with “.. relates to the possibility of adding 1” | Reviewer 2 |
| p.14 | Clarified logic of the current study, especially how measures were chosen, and clarified differences between Next Number and Highest Count as productivity measures | Reviewer 1 |
| **Methods** | | |
| p.15 (Sec 2.1) | Specified participant demographic pool | Editor |
| p.15, 17  (Sec 2.2.1) | Added flowchart of productivity decision tree (new Fig. 1) and referring text | Reviewer 3 |
| p.18  (Sec 2.2.1) | Added explanations for error cutoffs in determining Productivity | Reviewers 2 & 3 |
| **Results** | | |
| p.21, 25 | Renamed Section 3.1.1 and 3.1.2 headers for clarity | Reviewer 3 |
| p. | Updated Figure 2 (now Figure 3) to better visualize errors and show participant IDs | Reviewers 2 & 3 |
| p.25 (Sec 3.1.2) | Added text explaining analysis strategy for improved clarity | Reviewer 3 |
| p.23 (Sec 3.1.2) | Replaced “many Productive Counters (n=9)” to “some Productive Counters (n=9). | Reviewer 3 |
| p.26-27, 32-35. (Sec.3.1.2, 3.3) | Added Productivity by Initial Highest Count interaction to the analysis of Next Number accuracy and Infinity | Reviewer 2 |
| p. 31 (Sec. 3.2) | Added text clarifying comparisons between our work and Cheung et al (2017). | Reviewer 1 |
| p.32 (Sec. 3.2) | Referred to regression tables (Tables 3-5) earlier in the text for improved narrative | Reviewer 3 |
| **Discussion** |  |  |
| p. 39 | Clarified how children might think that successor function is bounded to a finite list | Reviewer 2 |
| p.31 |  | Reviewer 2 |
| p. 42-43 | Added discussion of study limitations and potential alternative explanations for productive counting, such as individual differences in domain-general cognitive abilities | Editor |
| p. 43 | Added reference to similarities in counting progressions observed in the educational literature (Frye et al, 2013) | Editor |
| **Supplemental Materials** |  |  |
| online | Added materials, data, and analysis code to the Open Science Framework: https://osf.io/z6ky3/ | Reviewers 2 & 3 |
| p.13-18  (Sec. D) | Added analyses for Next Number performance and Infinity regressions using stricter productivity classification (allow only 1 error, not 3), and referring text in main manuscript (p.22) | Reviewer 2 |

# Editor’s comments

**Thank you for submitting your paper for consideration at Cognitive Science. I have now received three reviews of your manuscript from consultants who are highly knowledgeable in the area of early number learning. I have also read it myself. We all agree that the study will be of interest to a general cognitive science audience. However, we do have some concerns and suggestions for improvements (appended below). Although I cannot accept the current version of the manuscript in Cognitive Science as is, I invite you to revise and resubmit based on the feedback here. I can’t be certain that you will be able to address the concerns in a revision, but I am optimistic.**

**The reviewers’ comments are appended in full below this letter. Please address each point made by the reviewers, especially Reviewer 3’s point about your method of choosing various error cutoffs and Reviewer 1 and 2’s concerns about the appropriateness of the explicit questioning method for determining understanding of the two aspects of infinity knowledge. Regarding Reviewer 1’s point, I would welcome various robustness checks in the supplementary materials. For example, what do the result look like if you only allow for one error instead of three?**

Thank you. We have now addressed these questions in our response below.

**In addition, please address these points:**

**1. Provide any additional demographic information on the participants in your sample. Do you have information on race/ethnicity, family income, parent education? I ask because on page 4 you describe the breakthrough of understanding number words happening around age 3.5 or 4 (in US English-speaking groups), but that’s only true for middle-income children in the US. Was the current sample of children predominantly from middle-income families?**

We have now provided general characteristics of the population from which we sampled in Section 2.1 (p. 13).

**2. Because Cognitive Science connects to several disciplines, including both psychology and education, it would be ideal if you could tie this work more to some of the early childhood education literature. For example, early childhood educators have charted what they see as a developmental progression in early number knowledge (see the IES Practice Guide on Teaching Math to Young Children), and that developmental progression places “number after” knowledge two steps before +1 knowledge. Could these ideas be acknowledged and incorporated into the lit review or discussion?**

Thank you for the suggestion; we now incorporate these ideas into the discussion (p. 41).

**3. Jennifer Kaminski has a highly relevant CogSci Proceedings paper that should be cited and discussed when presenting the previous work that has shown that successor function knowledge is acquired after CP-knowledge.** <https://mindmodeling.org/cogsci2015/papers/0186/paper0186.pdf>

Thank you for the citation, we have now incorporated it into our introduction.

**4. On page 9, you suggest two possibilities for the relation between how high a child can count and successor knowledge. However, there is also a third possibility. It’s possible that children’s non-numerical cognitive abilities such as pattern detection (apart from counting) could explain both. For example, children who are better at detecting patterns may more quickly become high counters and they also may be faster to develop successor knowledge. How does this possibility affect what can be concluded based on the current results? Perhaps a discussion of this issue would be helpful for readers.**

We have now included a discussion of domain-general factors and how they likely impact performance. We believe that learning to extract rules from data is a type of pattern detection, and would expect a domain-general test of pattern learning to correlate with counting ability. Also important would be tests of domain-general capacities that don’t relate directly to rule learning in content, but that might constrain learning - e.g., working memory. We have tested working memory capacity in another more recent study, and find that in that data set counting productivity measures have effects above and beyond working memory. We now incorporate these ideas into the discussion.

It’s important to note that domain-general differences between children likely can’t explain the pattern of findings we report as indicated by within-subjects comparisons. Our results point to dissociations between Highest Count and Productivity in predicting both knowledge of specific successors(on the Next Number task) and general successor function knowledge (on the Infinity interview). When controlling for Highest Count and age, Productivity uniquely predicted both children’s ability to identify successors for numbers outside their rote counting range, and children’s belief that numbers never end. Thus, at least some individual differences in conceptual knowledge about the successor function cannot be explained by the highest count measure. Consequently, we expect that any correlated non-numerical cognitive abilities would also not fully explain successor knowledge.

**5. An additional limitation that could be discussed further is that children could possibly be labeled as productive counters here without having knowledge of a productive counting rule. Some children store scripts in memory and may use those to succeed. Would it be possible for a child who relies heavily on scripts to be coded as a productive counter on your task? In other words, it’s possible that stating/reminding a child of the decade could activate memory of a verbatim script stored in memory, so they could start succeeding with that decade and the subsequent decade even though they don’t really understand the full rule. Thus, the present coding system may overestimate the number of children who truly have a productive rule. Perhaps creating a stricter criteria for coding children as productive counters would lead to stronger relations between that category and infinity knowledge?**

This is an interesting suggestion. In the paper we raise this possibility, and argue that it is unlikely given previously reported data and our data. On the view that children rely on rote memory, counting errors should occur at random locations in the count list, or, alternatively, should be more likely for less frequent number words than for more frequent numbers. Relying on scripts shouldn’t predict particular difficulty with decade labels, because they are, if anything, more frequent than words that come before or after them (e.g., “sixty” is more frequent than either “fifty-nine” or “sixty-one”). Therefore, we shouldn’t predict the attested finding that children often stop on decades (at 29, 39, 49, etc.) if they are simply relying on a memorized script. Also, assuming that children cannot recall “sixty” because they relied on a stored script, prompting of “sixty” should not be helpful for recalling “sixty-one” (not any more than prompting of “fifty-eight” should help with “fifty-nine”). Also, note that in our study, the Next Number task provides an alternative and potentially stricter test of Decade Productivity, as it does not have a rote counting component. We reasoned that if participants rely on memorized lists to succeed at the highest count task, then they should perform worse on the Next Number task for items outside their Initial Highest Count range than within it, since that is their best-memorized counting range. This was not the case (Section 3.1.2). Instead, as a group, children labeled as Productive Counters were equally successful for numbers within and outside their rote counting range. This suggests that children classified as Productive Counters did indeed have access to a productive rule for generating successors, as opposed to relying on stored sequences.

Perhaps a stricter test of productivity knowledge might involve unfamiliar or made-up numbers. For instance, a child should know that the successor of “a billion one” is “a billion two” and that the successor of “daxy-five” is “daxy-six”. We now include a summary of recent work testing this hypothesis in the Discussion.

# Reviewer #1:

**This manuscript reports one experiment designed to examine 3- to 6-year old children's explicit understanding of two important aspects of early number knowledge - the successor function and infinity. This is an important study because it addresses aspects of number knowledge that, in my opinion, are critical to the development of the number concept, but which have been understudied, and when they have been studied, often conflated with other aspects of number knowledge (e.g., cardinality). Indeed, most previous studies have focused on cardinal knowledge and acquisition of the cardinal principle as the critical transition point from an immature to a mature concept of number, and authors like Sue Carey have even claimed that acquiring the cardinal principle marks a conceptual shift whereby children move beyond their foundational, but limited, nonverbal representations to induce the structure of the natural number system. While often mentioned in published work, the successor function has not often been examined with empirical tests. Barbara Sarnecka's work (reviewed by the present authors) stands as an exception. Nonetheless, I feel like much more work is needed before we understand the developmental of this critical aspect of number knowledge and the role it plays in children's understanding of the natural numbers. The present study therefore makes an important contribution to the literature.**

**This manuscript is well-written and the introduction frames the work clearly. The sample is quite large, the appropriate ages are targeted. The analyses are appropriate. I feel that the authors were careful with their conclusions and that they did a nice job drawing conclusions based on the findings even though the data was not in line with their expectations.**

**My only real concern about the work is what seem to me potentially inaccurate definitions of the main constructs. Their measure for infinity seems pretty sound - the infinity interview questions 1-3 target the right construct, which is whether think the numbers ever end. Questions 4-6, however, do not seem to be getting at the successor function, but moreso other aspects of infinity. That is, understanding the successor function requires a child to understand (either implicitly or explicitly) that each number, n, has a unique successor that is generated by adding one to n. Asking children whether you can make a "biggest number" any bigger by adding to it (question 4), or whether you can "add one to the biggest number" (question 5), do not capture the notion of unique successor. Likewise, even question 6, whether it's possible to "keep adding one" doesn't really get at succession. Instead, it seems to be targeting infinity. I do believe that the "next number" task, though a strong test, is a good measure of children's successor knowledge because in order to succeed in that task kids must (implicitly or explicitly) understand how to generate the next number from the number given. Merely knowing that you can keep adding on to a number to make it bigger doesn't seem to me to capture the essence of the successor function. Indeed, the present data was equivocal about the relationship between highest count (and therefore decade productivity and by proxy understanding of the syntax of verbal counting) and successor knowledge based on their measure. I would like to see the authors make a convincing case of how their measure indexes true understanding of the successor function, rather than indexing just a different aspect of infinity.**

Thank you for these comments. In our revision, we are now more explicit regarding how we define these constructs and why. Mathematically, the successor function is actually quite narrow - it simply states that for any number, N, there exists an immediate successor. On this definition, a child who “knows” the successor function would only need to know that every number is followed by another number, without necessarily knowing the names of particular numbers. For example, an adult might know that an indefinitely large number like “a zillion” has a successor despite not knowing how to label it (or what the labels for much larger numbers are).

In order to deploy successor function knowledge in the Next Number task, the child needs to know specific number labels. But technically, a child might be able to label the next number for a finite list of words, but not know the successor function - i.e., that \*every\* number has a successor. For this reason it is important to ask children more general questions that are not tied to specific numbers, as we have in the Infinity Interview.

We have now explained this more explicitly in the paper (p. 8) and have cited various sources that define the successor function in the way we describe (e.g., Decock, 2008; Wright, 1983).

**The only other comment I had was whether asking children to explicitly discuss infinity and successor knowledge is really a good way to measure their understanding. I feel some adults might have trouble talking about the successor function, even though they clearly understand counting and how to generate the next number in a sequence. I guess I'm just wondering what the authors think is critical about asking children to demonstrate explicit knowledge of the successor function and infinity, rather than trying to devise an implicit measure (that differs from and improves upon the unit task, with it's limitations)?**

Thank you for the comment. In response to this comment, we have now added text to our Introduction and General Discussion.

We agree that this is the trickiest question raised by this work, and is a general problem for psychologists - how to assess construct validity and choose between alternative measures. And we hope that as this literature proceeds, researchers will derive ever-better measures.

The question raised by the reviewer ultimately boils down to two questions: (1) whether there exist better, alternative, measures that we know to have better validity, and (2) whether it is important that knowledge be probed using implicit vs. explicit measures. Relating to the first question, we reviewed the literature on infinity and successor function knowledge and selected this battery because it has been used repeatedly with similar results in multiple past studies, and because there are no better alternatives that we know of, or that we could devise. Note that all of this work finds that children reach ceiling on our measures by age 6 or 7 (e.g., Evans, 1983; Evans & Gelman, 1982; Hartnett & Gelman, 1998; Cheung et al 2017), suggesting that adults would have no difficulty with them. Also, these measures appear to be correlated with other numerically relevant tasks. Still, because we like to have converging evidence, we coupled the infinity battery with tests of how children reason about attested numbers, like the Next Number task. We do hope that other, better, measures might emerge, but currently there are few alternatives, and those that exist find much later competence, probably because they require much more sophisticated domain-general reasoning. For example, in Falk’s 2004 study, he creates a game in which the winner is the person who is able to name the largest number, and the dependent measure is whether the child chooses to go first, or to go second in the game; the inference is that children will always choose to go second if they understand that numbers can be infinitely generated, since the person who goes second can always use a rule to generate a larger number). Of course, there are many other demands of this task besides whether or not the child understands that numbers can be infinitely generated; for this reason, we elected to use the infinity battery presented in the present paper.

Regarding the question of explicitness, although we tested children with an interview that asks children to explicitly reason about numbers, we do not believe that answering such questions requires them to have explicit knowledge of the Peano axioms, per se. These axioms took humans thousands of years to explicitly derive after we first began using numerical symbols. However, the notion of infinity is much older in human history, and appears to be a basic intuition regarding the behavior of numbers that even our children seem to grasp: Success on the Infinity Interview does not require participants to spell out the successor function in terms of formal definitions like the Peano-Dedekind axioms, which we agree would be difficult for some adults.

Given that the development of these Infinity Interview responses is developmentally predictable across multiple studies, is related to other numeracy outcomes, and is direct in probing the underlying construct of interest, we believe that it is currently the best measure available, though we hold out hope that future studies will devise ever-better measures. We have now explained and justified our choice of tasks in the Introduction, so that future readers better understand our choice.

**Again, I think this work is interesting and important, I just wasn't sure if the piece I think is most critical - the successor function - was adequately measured. In light of this, I feel like the fact that highest count and next number performance were related suggests that implicit knowledge may be related to being a productive counter, but that this relationship just doesn't hold for explicit successor knowledge.**

# Reviewer #2:

**This study tests whether children's counting skills inform their understanding about infinity. Children performed two counting-related tasks and were questioned about their explicit knowledge about infinity. The results show that children who can count \*productively\* were more likely to believe that numbers are endless. The experiment was done in a thoughtful way and the paper is well written. Nevertheless, there are several major issues that question the significance and the validity of the findings.**

**There are many sentences that are difficult to make sense, difficult to understand, and even contradictory. It says "by learning that number words can be decomposed into decades and ones, children may realize that such rules can generate an infinite set of number words" (page 11). The rationale behind this logic is unclear. The highest count task tests whether children can generate the sequence of ones in the right order in any decade (note that the decades were prompted) until 99. That is, being able to correctly list from 1 to 9 after a given decade word is sufficient for a child to be categorized as a productive counter. How does that knowledge precisely let them realize the infinite nature of numbers? As an example, imagine that children are learning the sequence that starts with A1, A2, … A9, then B1, B2, … B9, all the way until Z1, Z2, … Z9 (i.e., it's a finite set). Even though children may master the whole list, this is a finite set and there is no pressure for children to think that the list must go on. Again, it's unclear how the authors have come to hypothesize that such kind of bounded sequential knowledge triggers the concept of infinity. A clearer logic behind this hypothesis is needed.**

The intuition behind our hypothesis is that when children acquire morphologically complex words (e.g., walk-ed; chair-s) they can either learn words item-by-item or they can acquire rules that permit the generation of words (e.g., past tense; plural), which in theory can generate an unbounded number of forms. Children appear to favor the adoption of rules when the number of cases that can be described by a rule is large - e.g., an open class - as in the cases of the past tense and plural. Our hypothesis is that after having learned 2-3 decades in an item-based way, children derive a rule to describe the combination of “decade+unit”, and this rule generates novel cases, effectively guaranteeing that numbers can be generated indefinitely (and thus that they are infinite).

We have now described this in further detail in the Introduction (p. 8-12).

**It is difficult to understand the interpretation that "the [successor] function is bounded to a finite list" (p.36). In fact, this is contradictory, considering that the paper is centered on the Peano-Dedekind Axioms (1), according to which the successor function "generates an infinite set" (p.6). Such an expression should be removed.**

Thank you for the comment. We agree with the reviewer that the successor function does generate an infinite set. We have edited the sentence to clarify that we are discussing children’s developing knowledge as opposed to the formal successor function.

**The way that productive counters are categorized seem ad hoc and is complicated. There seems to be no rationale behind certain decisions. For example, why allow 3 errors? Why up to 2 decades?**

Thank you for the suggestion; we have now explained the choice of error cut-offs in greater detail (Methods, p.17), as we realize it does seem arbitrary without further context. While the application of 3-error/2-decade productivity criteria was not specifically designed with this study in mind, it was drawn from previously published, pre-registered reports in which this measure was shown to be significantly related to other measures of counting productivity (Schneider et al, 2020). The 3 error criterion was chosen to allow for errors at decade transitions, which are often irregular, as well as 1 additional error per decade. The 2 decade criterion was chosen both to provide evidence that children’s counting after an error is rule-governed, which is more difficult to determine with a 1 decade limit, and also to accommodate children who were able to count to the maximum number possible in the Highest Count task (100). Note that in this work, we applied this criterion to children who made errors in the count list that were not corrected (i.e., mid-decade errors, or errors at decades that were missed by the experimenter), but were still able to recover from those errors. We acknowledge that other reasonable criteria could surely be adopted, too, and that this work is in large part exploratory. For this reason we will share our data on the OSF for others to analyze and subject to alternative criteria. Critically, however, our criteria were specific prior to data analysis, and based on our best guess regarding what a productive rule should predict.

In addition, we have added a section in the Supplemental Information (Section D) where all analyses were repeated using a more conservative criterion that allows for only one error. This resulted in 6 participants re-classified as Non-Productive counters (5% of our total sample). These analyses led to the same conclusions as reported in the main text regarding how Productivity influences Highest Count performance, Next Number performance, and Successor Knowledge. However, there was a different result for regression analyses predicting Endless knowledge and Full knowledge of infinity. First, the original findings reported that Productivity group significantly predicted Endless knowledge; this effect did not replicate using the stricter definition. Second, the original findings did not yield any significant predictors of Full infinity knowledge when controlling for age; however, the stricter Productivity definition significantly predicts Full infinity knowledge, even when controlling for age or Initial Highest Count. Given that Full infinity knowledge requires Endless knowledge, these results generally converge.

In summary, these analyses indicate three robust findings: (1) there exist significant individual differences in knowledge of the decade+unit rule, which affect performance on counting and Next Number tasks; (2) knowledge of this productive rule does not predict successor function knowledge (i.e. the belief that we can always add one); (3) Productivity does predict infinity understanding (i.e. the belief that numbers never end), although particular regression results may depend on how strictly we define productivity and infinity knowledge (i.e. Endless knowledge vs. Endless+Successor knowledge)

**Figure 2 is quite confusing as well. The 6th-10th children from the left in Fig 2a have long orange lines, indicating that they made errors across eight to nine numbers. Why are they still considered productive, if they made more than 3 errors? I understand that it's hard to quantify these rather open ended responses. However, showing the robustness of the results against some of these decision parameters (and the replicability/reliability of these categorization procedure\*) must be considered.**

**\* This should be beyond the scope of this paper, though.**

Thank you for the comment; we have updated the figure (now labeled Fig. 3) for clearer representation of errors. The orange lines represent continuous skipped sequences, which are coded as a single error. The specific participants you mention (participants 46 to 50 in Fig 3a) had skipped the continuous thirty-X sequence, counting: “28, 29, 40, 41, …”. To better indicate that these are a single error, we now use grey lines to represent skipped sequences. Thus, orange dots indicate either individual errors (e.g. “12, 13, 18, 15, 16”) or the beginning of a missed sequence.

In this way, the number of orange dots represent the total number of errors made. Participant 47 is categorized as a Productive Counter according to the improvements they made after receiving Decade Prompts: they counted past their Decade-Change Error at 49 by at least 2 decades after being prompted with ‘50’ and ‘60’.

Our shared dataset includes children’s raw counting sequences with any experimenter prompts. We hope this allows others to explore alternative criteria for classifying children’s responses.

**The term "syntax" in the title is misleading because the measure of productivity is rather related to sequence. In fact, the authors appropriately explain it in terms of the "verbal count sequence" in the abstract, and that phrase is more appropriate in the title as well.**

We understand why this use of the word syntax might be confusing, and so we’ve now made edits to make our usage clearer. Central to this, our paper is interested in how the words within the counting sequence are generated, rather than how children learn relations between words within the sequence. Our hypothesis is that count words are generated by a syntactic (or morphological) rule. We now take greater care to explain this rule in the paper, and how it relates to infinity. Note that the way we use the words morphology and syntax to refer to counting structures is compatible with how it is used elsewhere in the literature (e.g., Rule et al., 2015; Yang, 2017; Comrie, 2011).

**There are issues with the results from the infinity questionnaire. First, the results do not replicate their earlier study: "Prior research suggests that how high children can count is related to their ability to identify successor relations for known numbers" (p.33). Unfortunately, IHC did not predict Successor knowledge, Endless knowledge, or both combined (Tables 2-4) in this study. These results question the reliability of the test. In fact, the authors themselves acknowledge that children's responses to these questions are extremely unreliable (p.36). Second, it's hard to understand how some children (24%) are only successor knowers and other children (9%) are only endless knowers. Logically speaking, successor knowledge should lead into the concept of endlessness, which leads me to think that 24% of the children ended up being categorized that way because the questionnaire does not truly tap into their understanding of successor or endlessness, or at best provides a very unreliable measure of such knowledge.**

Thank you for the comment and for pointing out the two sources of potential confusion. It is important here to clarify what is meant by “successor knowledge” in our study and what is measured in other studies. In most other studies (e.g. Cheung et al, 2017 and Sarnecka & Carey, 2008), knowledge of successor relations is measured by the Unit Task, which asks children to identify number words that represent a quantity one larger than a target set. Thus, the “ability to identify successor relations for known numbers” (p.36) involves knowing that adding one object to a set labeled “four” now elicits the label “five”. In contrast, in our study Successor Knowledge on the Infinity Interview (Tables 3-5) refers to endorsing the claim “you can *always* add one to a number”. For improved clarity, we updated our manuscript to refer to this as “successor function knowledge”, to contrast with item-based knowledge about particular successors.

Due to the difference in defining “successor knowledge” between our study and Cheung et al (2017), we believe that our results are not a failure to replicate. Instead, our results are compatible with Cheung et al (2017) for two reasons. First, as described above, “Successor Knowledge of Infinity” refers to knowledge that “you can always add one”, i.e. successors exist for all numbers, without requiring children to verbally identify them. Thus, we think that it is a different construct than the knowledge of specific successor relations tested in Cheung et al on the Unit Task. In fact, results from Cheung et al (2017) support this distinction: “children’s performance on the Inﬁnity Task remains relatively poor until they are near ceiling on the [Unit] Task.” (Cheung et al, 2017, p.29). Second, while Cheung et al (2017) find that children with full Infinity Knowledge (i.e., both Endless & Successor beliefs) tend to have higher Initial Higher Counts, they did not explicitly test the relationship between IHC and performance on the Infinity Interview while controlling for age. We also find that children with full infinity knowledge have higher IHC (M=68) than other participants (M=46), however the relationship between IHC and Infinity knowledge is not significant after controlling for age (Table 5).

Thus, we believe that our findings do not reflect a failure to replicate and instead reveal additional detail about the relationship between counting and infinity understanding.

The second issue refers to the existence of children who we can label “Successor-only knowers of Infinity” and “Endless-only knowers of infinity”. You are right to point out that the claim “you can always add one to a number” (i.e. Successor Knowledge of Infinity) logically implies the claim “numbers never end / there is no biggest number” (i.e. Endless Knowledge of Infinity). In fact, in the paper we acknowledge this logical implication in our discussion of the Peano-Dedekind axioms (p. 11). While this might reflect unreliability in the Infinity Interview, another interpretation would be that 4- to 5-year-old children have not fully recognized the logical implications of the “Successor Knowledge of Infinity” for a fully mathematical understanding of the natural numbers. Very generally, much of mathematics is an exploration of the entailments of first principles, little of which is known automatically, but must be discovered and proved by mathematicians. Support for the idea that this task is likely reliable, and that children learn the entailments of the successor function gradually comes from the compatibility between our findings and previous studies using similar Infinity interviews, including Cheung et al (2017) and Hartnett & Gelman (1998). For instance, Cheung et al (2017) found 33 “Successor-only knowers of Infinity” among a sample of 100 4-7-year-old children. Similarly, Hartnett & Gelman (1998) classified about 50% of children aged 5-6.5 years as “Waverers”, which correspond to children with either “Successor-only” or “Endless-only” knowledge of infinity but not both. Finally, both these papers find that older participants were often reliably classified as Full Infinity Knowers, suggesting that variability in Infinity Knowledge status comes from variability in participants’ knowledge of infinity concepts.

We have now added text to the paper that clarifies how our work relates to previous reports (to address questions related to replication), and have also noted that although a recursive successor function entails infinity, it is reasonable to expect that children do not automatically compute the entailments of their beliefs.

**There is a high correlation between IHC and Productivity Group, even if those with IHC=99 are removed. For example, the majority of the Productive Counters have IHC>=40; the majority of the Non-Productive Counters have IHC<40. Therefore, in their regression analyses (e.g., Tables 2-4), IHC must be entered together with Productivity Group (and the interaction should be assessed as well). Otherwise, it is not possible to tell whether the effect of Productivity Group is due to productivity only or due to productivity combined with IHC.**

Thank you for the suggestion. We are also interested and have updated the results section (Section 3.3) to include this logic. Because the only models that yielded a significant effect of Productivity Group was in predicting Endless knowledge, we have updated Table 4 with additional models controlling for IHC and the Productivity:IHC interaction. These additional regression models yielded the same conclusion; Productivity was a significant predictor of Endless knowledge even when controlling for IHC.

We have also included a Productivity by IHC interaction in predicting Next Number accuracy (Section 3.1.2). These results are now updated.

**Minor issues:**

**On page 11, it is said "[children may] infer that numbers must be infinite, even without yet understanding how this recursive rule relates to cardinality" when raising an alternative hypothesis to the idea that successor knowledge must mediate the relationship between counting knowledge and infinity knowledge. That statement is a possibility, but none of the tasks tests cardinal knowledge, so that sentence seems out of place and needs to be revised if not removed.**

Thank you for the suggestion; we have now replaced the word “cardinality” with “the possibility of adding 1” (p. 12)

**Why are 40 and 70 categorized as a mid-decade number? The step from 40 to 41 is vastly different from the same steps in between 41 and 49.**

Our hypothesis was that lacking knowledge of decade words (“forty”, “fifty”) might mask children’s latent knowledge of the decade+unit rule for composing two-digit numbers. Thus, we classified numbers as either cross-decade (child has to generate the correct decade word, e.g. 29 >> 30) or mid-decade (correct decade word is already provided). Since the step from 40 to 41 does not require the child to generate the decade word “forty”, this trial counts as a mid-decade number.

# Reviewer #3:

**The authors report a study with 4 and 5 year olds, investigating their knowledge of counting (through 3 tasks) and two aspects of "infinity" (successor and endless numbers). They investigate the relations among these measures to shed light on which skills may be more/less associated or accessing distinct aspects of the concepts. The pattern of results in terms of children's performance on the three different counting tasks is interesting in and of itself, but the primary finding (or, what I take as the author's primary finding) is that they do find evidence of distinct relations between productive counting and successor vs. never-ending numbers — suggesting that these two aspects of infinity are differentially related to the counting structure knowledge.**

**The inclusion of these tasks in the same children provides important theoretical and methodological contribution to the study of children's number knowledge. However, the are aspects of the manuscript that are difficult to follow or otherwise lower the potential impact. My more detailed comments are below:**

**Method**

**In general the method was clear, although the coding decision-making sections were sometimes difficult to follow. I wonder whether a decision tree (even just in supplemental) would be helpful to more clearly communicate how these bins were determined.**

Thank you for the suggestion; we have added a decision tree (Fig. 1) to the supplemental information to describe the Highest Count coding scheme.

**Also, can you provide more information/justification for the 3 errors cut off? Why 3 or fewer vs. 4 or more errors? What was the distribution of error numbers? (i.e., did it tend to be 1 or 2 vs. a lot? or were there many children at those border cases of 3 vs. 4 errors, which seems rather arbitrary?)**

Thank you for the comment. We have now added details to the paper, and also describe our logic above in response to Reviewer 2’s question above, which is based on the logic of previous, pre-registered analyses.

**- Also, of the children that made lots of errors, did they tend to be the same error (e.g., always says 5, 4, 6 — 25, 24, 26 — 35, 34, 36 etc.) — because if so, this would result in 9 errors, but seems unfair to call them not productive counters.**

**- More descriptive information about these categories, the number of / distribution of children within the categories for various reasons (i.e., because the made no errors vs. 1 error etc), would help clarify and provide some justification for why 3 errors was used as the cut-off**

**- Actually, now having read further, Figure 2b can speak to this already a little bit, so it may be helpful to refer to this figure to discuss whether there were any/many cases that were given an "arbitrary" category, e.g., 4 errors vs. 3 errors. Numbering the participants to refer to specific borderline cases or which cases went into the given category for which reasons could be helpful for understanding the justification of this cut off.**

Thank you for the suggestion; We have added participant indexes to Fig 2 (now renamed Fig 3).

In regards to the distribution of children’s errors, we did not find any children who made the same error that repeated across decades. Children who made many errors tended to recite numbers in random sequence. Participants’ raw count sequences can be found in our published dataset, available on the OSF repository.

In addition, most participants made 3 or fewer errors; only one participant made 4 errors. For participants who counted to 100 on their own, only one participant made 4 errors (Fig 3b., subject 23). Participants who were classified as unproductive either stopped counting on their own in the middle of a decade, or stopped counting after failing to continue counting past the first or second decade prompt (n=18).

**Results:**

**In general, the results are difficult to follow, but the figures and tables greatly help the reader. It would be helpful to future subset the results beyond just the task being analyzed (especially since section 3.1.2 Number Number Task, for example, also includes substantial analysis involving the counting tasks as well). For example, organized by research question or categories like "Describing Performance" vs. "Relations Among Tasks" would help the reader navigate the substantial information being provided.**

We have renamed section headings: “3.1.2: Productivity and Next Number performance”

**Page 19: "However, many Productive Counters (n=9) had an Initial Highest Count of 29 or lower" It's not clear that 9/73 productive counters would be considered "many". It's an important note that there are some Productive Counters with low IHCs, but Given the subjective interpretation of "many", it may be better to simply say "some"**

Thank you for the comment; we have edited that sentence for clarity.

**- Section 3.1.2: first analysis is a t-test with all children, second is a mixed effects regression with a subset of children —> it's a little hard to follow why the change in analytic approach. The explanation for why looking at a subset of children is clearly provided, but the motivation for changing the analytic strategy is unclear. I think this might just be an issue of more fully explaining the rationale, but it may also need some alignment of analytic approaches**

The general analytic approach in this paper is the use of generalized linear (mixed effects) models. We have edited the text in Section 3.1.2 to explain that the first t-test analysis provides only a descriptive comparison of productive and non-productive counters without controlling for covariates of interest (e.g. age and initial highest count).

**- The paragraph of discussion of this section on Page 23 is also very hard to follow, it would be helpful to provide a bit more information about what analyses or aspects of the methods these claims are following from (e.g., the idea that productive counters could have only been tested on small numbers was jarring, until I realized that was because they could have erred on 20, and then gone as far as 40 without another error; but it would be helpful to spell that out more in this section; the "first analysis" isn't clear - do you mean the t-test comparing Productive and Not Productive? If so, it's unclear how that suggests that Next Number and Productivity capture different aspects of counting knowledge.**

Thank you for the comment; we have edited the text so that Section 3.1.2 clearly describes 3 separate analyses (t-test, regression controlling for IHC and age, regression including specific items). We have also edited the discussion paragraph (page 37) to refer to these specific analyses and explain which claims follow from which analyses.

**- it would be useful to point to tables 2-4 earlier in the results section, as it's much easier to follow the narrative explanation of the results along with the tables; for e.g., indicating that results are in Table 2 when stating: "Interestingly, in predicting children's Successor Knowledge, none of the three predictors explained a significant proportion of additional variance compared to the base model."**

Thank you for the suggestion; we have referred to these tables earlier in the results section (p.32). We also refer to the relevant table when describing each set of regression analyses by outcome variable (Successor, Endless, or Full Infinity knowledge).

**Other Major Comment:**

**I highly recommend the authors make their data, analysis code, and/or materials available on a repository (e.g., the open science framework, although there are also others) so that others can easily replicate and reproduce their results. Notably, this would be particularly useful given my question about the categorization of individuals into productivity bins - if others are unsure of this categorization, they can more closely inspect or re-analyses with different decisions, allowing others to investigate the robustness of these results in other ways**

We have made our data, analysis code, and experimental protocol available on the Open Science Framework: <https://osf.io/z6ky3/>