How Reliable is the Give-a-Number task?

Elisabeth Marchand (emarchan@ucsd.edu)
David Barner (dbarner@ucsd.edu)

University of California, San Diego
Department of Psychology
9500 Gilman Drive, La Jolla, CA 92093-0109, USA

Abstract

The Give-a-Number task has become a gold standard of children's number word comprehension and has been increasingly used to organize debate in developmental psychology. In this task, the experimenter asks children to give specific numbers of objects (e.g., 1 to 6), and based on their pattern of responses, children are classified into stages that can be readily related to other developmental milestones. The increasing popularity of Give-a-Number raises the question of how reliable it is, since the size of a correlation between two different tasks cannot reliably exceed the test-retest reliability of either measure taken individually. In Experiment 1, 2- to 4-year-old children were tested twice in a single session with Wynn's (1992) version of the Give-a-Number task, which features a titrated design. In Experiment 2, we tested a second group of children with an alternative version that uses a larger number of trials in a non-titrated design. We found that in both cases the task was highly reliable in differentiating children who could accurately count from those who could not, but that reliability differed for specific numbers, and was more reliable for very small numbers (i.e., "one" and "two") than for slightly larger ones (i.e., "three" and "four"). We discuss practical implications of these results for researchers studying numeracy and discuss further directions to assess the validity of the task.

Keywords: Give-Number task; concordance; number acquisition

Introduction

Preschool children are often good at reciting the count list, but, early in development exhibit surprisingly little understanding of number word meanings and how to accurately count sets. Over the past 40 years, a large corpus of studies in the field of number cognition has revealed that children acquire the meanings of number words in highly protracted stage-like sequence, and that this basic pattern is present across a range of different cultures and language groups. In the U.S., English-speaking children typically begin by learning the count list at around the age of 2 as though it were a single expression ("onetwothreefour…"), without attaching meaning to the individual words (Carey & Sarnecka, 2008; Fuson, 1988). For this reason, these children are often called "non-knowers". Not long after, children learn the meaning of their language's word for "one", which means, in practice, that they can provide one object upon request but can't reliably give accurate amounts for larger numbers. Over the course of many months, children then learn the meanings of "two", "three", and "four" in sequence. This is followed by a form of breakthrough in which children learn to use their memorized counting routine to label and generate any set within the range of their count list (called the Cardinal Principle Knower or CP-Knower stage). Data compatible with this basic pattern have been documented in English, French, Spanish, Japanese, Russian, Slovenian, and Tsimane amongst others (Almoammer et al., 2013; Piantadosi, Jara-Ettinger, & Gibson, 2014; Sarnecka, Kamenskaya, Yamana, Ogura, & Yudovina, 2007; Wagner, Kimura, Cheung, & Barner, 2015).

This apparently robust developmental sequence has been demonstrated in large part by using the Give-a-Number task (Give-N). In this task, children are presented with a set of objects (e.g., 10 apples), and are asked to "give" subsets of this set (e.g., by placing them into a container), often starting with one – e.g., "Can you put *one* apple in the plate?". Though versions of the task were used as early as the 1970s (Schaeffer, Eggleston & Scott, 1974), Give-N became a type of gold standard to study number word comprehension following its use by Wynn in two papers (Wynn, 1990, 1992), which have now been cited nearly 2000 times. In one version of this task, used by Wynn, the trial structure of Give-N is titrated, such that if a child responds correctly to a request (e.g., giving exactly 2 objects when asked for *two*), they are then tested with the next largest number (e.g., *three*), whereas if they fail they are tested on a smaller number. This procedure is then repeated until the experimenter is able to identify the highest number that a child can succeed at, 2 out of 3 times (i.e., knower level; see method for more details). Other studies have used an alternative, non-titrated, version of the task in which children are tested on all numbers of interest (e.g., 1, 2, 3, 4, 5, 6, 8, 10) three times each in pseudo-random order. The reason why some studies favor one version over the other seems be to hypothesis driven; for example, studies interested in specific numbers (e.g., how "one" or "two" are acquired; Almoammer et al., 2013; Sarnecka et al., 2007) may use the non-titrated version of Give-N as it ensures that all numbers of interest (e.g., one, two and three) will be tested at least 3 times, unlike with the titrated version. Importantly, past studies generally assume that different versions of Give-N are interchangeable as diagnostics of knower level.

Using this framework, numerous studies have begun to ask how these stages of number word development are related to other developmental measures, such as vocabulary size (Negen & Sarnecka, 2012), comprehension of grammatical number (Almoammer et al., 2013; Le Corre, Li, Huang, Jia, & Carey, 2016; Sarnecka et al., 2007), or later mathematical achievement (Chu, vanMarle, & Geary, 2016; Purpura & Simms, 2018). Critically, however, the replicability of the overall knower level framework does not itself assure the reliability of individual knower levels, and therefore doesn't guarantee that testing correlations between knower levels and other factors will generate interpretable results.

Currently, the reliability of the knower level status of any particular child within a dataset is not known. This is important because the strength of a reliable correlation between two observations (e.g., knower level and vocabulary size), *r*(ObservedA,ObservedB), is bounded by both the size of the correlation between the true value of the variables being measured, *r*(TrueA,TrueB), and the test-retest reliability of these measures taken individually, reliabilityA, reliabilityB (Nunnally, 1970). Thus, as noted by Vul, Harris, Winkielman and Pashler (2009), in a scenario in which a true correlation between two variables is 100% but the test-retest reliability is .7 for one and .8 for the second, the highest reliable correlation that can be detected is .75 (i.e., 1 x √(.7 x .8) = .75). In the current context, this means that if individual knower levels – e.g., the 1-knower stage – exhibit low reliability (e.g., .3), then the size of expected correlations between knower level and other variables should also be low. Also, it means that a particular knower level assignment might overestimate – or underestimate – a child's true knowledge. More generally, interpreting correlations between knower levels and other outcomes hinges critically on the reliability of the Give-N task.

In the present study, we investigated the reliability of the Give-N task in two studies. In Experiment 1, we assessed the test-retest reliability of Wynn's titrated version of Give-N and in Experiment 2, we measured the test-retest reliability of the alternative non-titrated version, which we expected might offer stronger reliability than the titrated version, because it features more trials and tests children using the same trial structure across administrations. Aside from these two different methodologies, there are also other ways in which the administration of Give-N likely differs across labs that could affect the reliability of the task. Here, we systematically assessed the potential impact of one such factor, testing location.[1] Specifically, in both experiments, we tested children across different settings (within subjects) – either in lab or outside of lab (i.e., museum, preschool) – to assess the impact of experimental environment on knower level reliability.

Experiment 1

Method

Participants In total, 81 English-speaking children, aged 2;2 to 4;1-year-old were included in the study (*M* = 3;4 years). This age range was targeted as previous studies have shown variability in children's knower-levels at this age. An additional 35 children were excluded from analysis because of failure to complete all 3 tasks (n=11), being outside the targeted age range (n=4), because English was not their primary language, because of language delay (n=3), or experimenter error (n=17). Participants were recruited from a parent database (lab), preschools and museums in San Diego. Informed consent was obtained from the parents. The study received approval by the ethics committee of the University of California, San Diego.

Materials and procedure In order to assess the influence of testing location, children were tested either in the lab or offsite (i.e., preschool and museum). The testing environment in the preschools and in museums was similar and consisted of a relatively quiet corner of a room made available by the staff. In the lab, the testing environment was more quiet than off-site and possible distractions were limited (i.e., proximity of games, toys and noise). Each session lasted approximately 8 min and included (1) Give-a-Number task 1, (2) Highest Count task and (3) Give-a-Number task 2. All participants were administered the tasks in this order. Children received a small prize for their participation at the end of the session.

Give-a-Number Task (Titrated) This task was adapted from Wynn (1992). Stimuli consisted of a puppet, a plastic plate, and a pile of small plastic toys. Children were asked to provide a certain number of toys in the following way: "*Mr. Monkey is very hungry. This is a plate and these are your bananas. I want you to put bananas in the plate for Mr. Monkey ok? Listen carefully! Can you put N banana(s) in the plate?* (N is the number word). *Put N banana(s) in the plate and tell me when you're all done*". After this first prompt, children were asked to count to verify that they had provided N (i.e., *"Is that N? Can you count and make sure?"*), and if they chose to fix their answers only their final responses were recorded. Children were always asked for 1 first and then 2. If the child succeeded on both trials, the experimenter then asked for 3, otherwise, they asked for 1. The next requests depended on the child's pattern of response: if the child succeeded, the experimenter asked for N+1 and if the child failed, they asked for N-1. The lowest request was 1 and the highest was 6. Consistent with Wynn (1990)'s criteria, children were credited as N-knowers (e.g., two-knowers) if they correctly gave N objects at least 67% of the time when asked for N, and failed to give the correct N at least 67% of the time at a request for N+1. In addition, for the child to be credited as an N-knower, 2/3 of their responses of N objects had to be in response to requests for N (e.g., such that a child who gives 2 objects across all trials

[1] Various others exist. For example, labs test different numbers, provide children with different numbers of objects, order of trials, type of follow-up questions children are asked, number of objects presented and environment in which children are being tested.

would not be credited with knowing the meaning of *two*). Finally, children were credited as CP-knowers if either they succeeded on 67% of trials for 5 and 6 or responded correctly to each request, 1 to 6, consecutively. Finally, except in this last instance (of CP-knowers), children were tested with a minimum of 2 trials for N in order to verify that they were an N-knower.

Highest Count Task (HC) This task had two goals: first, to serve as a proxy for exposure to numeracy in our model comparison of the two Give-N tasks (titrated vs non-titrated), and second, as a filler task between Give-N tests. Participants were asked to count as high as they could. The last number reached before stopping or making an error was recorded as the child's highest count.

Results & Discussion

Give-a-Number Table 1 shows the distribution of knower-levels in the first and second assessment of the titrated Give-N task. We first assessed the agreement and reliability of the task by including all knower-levels (0 to CP) in a 7x7 contingency table (see Figure 1 for example of contingency table). Reliability was measured using the weighted Kappa statistic (Cohen, 1960).[2] We obtained an agreement of 77% and a Kappa of 0.866 (unweighted 0.709), which corresponds to what previous studied classify as "excellent" reliability (Landis & Koch, 1977; but see Sim & Wright, 2005, for disagreement regarding how to describe different levels of reliability). However, as indicated by Figure 1, the rate of effective agreement (in percentage) across different knower levels was quite variable, ranging from 18% to 76%. These first two results suggest that when all knower levels are considered together, the Give-N titrated task has a high degree of reliability, but that individual knower levels differ substantially and may not be uniformly strong predictors in statistical tests.

To explore this issue further, we calculated agreement and Cohen-Kappa for subset-knowers, non-knowers and CP-knowers separately. For the subset-knower analysis, we created a 6x6 contingency table with the knower-levels 1 to 5, as well as a new category of non-subset-knowers (binning together non-knowers and CP) for Give-N Test 1 (T1) and Give-N Test 2 (T2). We first calculated the reliability of knower levels within the subset range, taken together, and found an agreement of 63% and an unweighted kappa of 0.714, which is considered "substantial. Next, for the non-knower analysis, we generated a 2x2 contingency table with non-knowers and non-non-knowers (i.e., all subset knowers and CP-knowers) at Give-N T1 and T2. We obtained an effective agreement of 80% and a reliability of 0.951. Next, for the CP-knower analysis (CP vs non-CP at T1 and T2) we found an agreement of 76% and a reliability of 0.827, which is considered excellent. These last two results suggest that the non-knower and CP-knower classifications are highly reliable, and somewhat more reliable than classifications within the subset stage, when all subset knower levels are considered together (though as already noted, reliability within the subset stage varies between individual knower levels, as shown in Figure 1).

In some past studies (e.g., Sarnecka & Carey, 2008), researchers have been less interested in whether a child is a specific N-knower (e.g., one-knower), and more interested in whether they are a CP-knower or instead have not yet learned to count accurately, and are a subset knower or non-knower. Relatedly, many studies simply lack the power to analyze individual knower levels as predictors. In our next analyses, we therefore asked whether a child classified as, for example, a subset-knower at Time 1, was likely to be a subset-knower again at Time 2. To do this, we divided knower-levels in 3 groups: non-knowers, subset-knowers (1K to 5K) and CP-knowers.[3] We then created a 3x3 contingency table with knower-level groups at T1 and T2. Here, we found an overall agreement of 89%, and a weighted Kappa of 0.873, which is considered excellent. This suggests that children who were classified as subset-knowers in the first assessment are very likely to remain subset-knowers in the second assessment, as are non-knowers and CP-knowers.

We next asked whether, when discrepancies existed between knower-levels within a subject, knower level systematically increased or decreased between Time 1 and Time 2. An increase could signal a practice effect while a decrease would suggest a fatigue effect. In total, more children exhibited a decrease in their knower-level from Give-N 1 to Give-N 2 (decreased n=13; increase n=6) but this difference was not significant (p=0.11). In addition, most of these children had knower-levels that differed by one level (difference of 1 level, n=11; difference of 2, n=8), though again this difference was not reliable (p=0.17).

Table 1: Distribution of Knower-Levels at the first (T1) and second (T2) assessment of Give-N titrated[4]

	0K	1K	2K	3K	4K	5K	CP
T1	9	14	16	5	7	3	27
T2	9	15	15	8	6	4	24

[2] The overall agreement corresponds to the total number of matches between the first and second assessment of Give-N titrated divided by the total number of observations. The effective agreement is the number of matches divided by the number of observations that include at least one of the Knower Levels in consideration. However, both values are inflated indexes of reliability as they don't consider the agreements that could have occurred by chance. Kappa is considered to be an improvement over % agreement as it controls for chance. Also, weighted kappa is considered to be more appropriate for ordinal scales, as it attaches greater weight to large differences between ratings than to small differences.

[3] In task 1, there were 9 children classified as non-knowers, 45 subset-knowers (1K to 5K) and 27 CP-knowers. In task 2, there were 9 non-knowers, 48 subset-knowers and 24 CP-knowers.

[4] Here, 0K refers to non-knowers, 1K to one-knower, 2K to two-knower, 3K to three-knower, etc., and CP to cardinal-principle-knower.

Testing Location We found no difference in agreement between knower-levels depending on the testing location (in lab vs offsite; p=0.29)

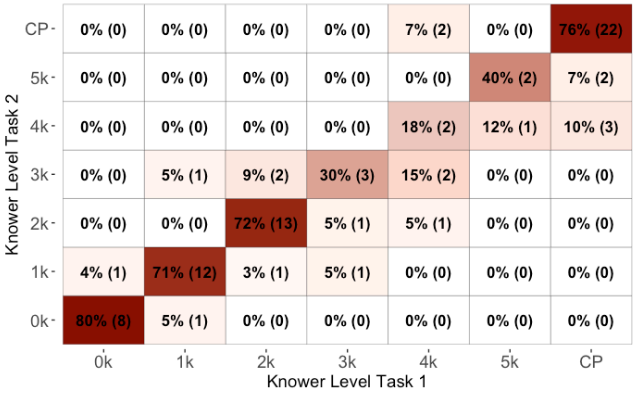

Figure 1: Knower-level classification in the first (T1; x axis) and second assessment (T2; y axis) of Give-N titrated. The percentages represent the percent effective agreement – i.e., the agreement calculated over not all paired knower-levels, but those paired knower-levels in which at least one belongs to the knower-level in consideration. The number in parenthesis represent the frequency of the paired knower-level. The color scale is based on the proportion of effective agreement, where darker red represent higher agreement.

Experiment 2

In Experiment 2, we assessed the test-retest reliability of the non-titrated version of Give-N.

Method

Participants In total, 81 English-speaking children, aged 2;6 to 4;0-year-old were included in the study (*M* = 3;4 years). An additional 20 children were excluded from analysis because of failure to complete all 3 tasks (n=12), being outside the targeted age range (n=5), because English was not their primary language (n=1), or experimenter error (n=2). Children were recruited in the same way as in Experiment 1. The study received approval by the ethics committee of the University of California, San Diego.

Materials and procedure The testing environments were identical to Experiment 1, except that children were presented with a non-titrated version of Give-N, twice.

Give-a-Number Task (Non-Titrated) This task was identical to the titrated version used in Experiment 1, aside from the trial structure. In this task, each child was given 15 trials: three trials for each of the numbers 1, 2, 3, 4, 6. We created two lists of trials in a pseudorandom order. All children were presented with both lists, either at Time 1 or Time 2 and we counterbalanced which list came first across children. Note that since we did not ask for 5, children could not be classified as 5-knowers in this version, unlike in the titrated task (note, however, in Experiment 1, only 5 children were ever classified as a 5-knower). The criteria to assign knower-level were the same as those used in the titrated version: children needed to correctly give N two out of three times when asked for N, and fail to give the correct N two out of three times for N+1. Again, children could not use N more than 50% of the time for requests other than N and children were credited as CP knowers if they could correctly give six, two out of three times.

Highest Count Task (HC). The task was identical to Experiment 1.

Results & Discussion

Give-a-Number Table 2 shows the distribution of knower-levels in the first and second assessment of the non-titrated Give-N task. We first calculated agreement and Cohen's Kappa including all knower-levels (0 to CP) in a 6x6 contingency table. We found an agreement of 73% and a weighted Kappa of 0.815 (unweighted 0.650), which corresponds to excellent reliability. The contingency table in Figure 2 illustrates children's knower-levels in the two iterations of the non-titrated task as well as their agreement.

Next, as in Experiment 1, we explored the reliability for subset-knowers, non-knowers and CP-knowers separately. For the subset-knower analysis, we created a 5x5 contingency table with knower-levels 1 to 4 and a non-subset-knower category for Give-N T1 T2. We found an agreement of 58% and an unweighted Kappa of 0.661, which is considered substantial. In the non-knower analysis (2x2 contingency table), we obtained an agreement of 57% and a reliability of 0.926. In the CP-knower analysis, we found an agreement of 76% and a reliability of 0.803. Similar to Experiment 1, these results suggest that reliability is affected by knower-levels such that the reliability within the subset-knower level is lower than that of CP-knowers and non-knowers.

Next, we assessed the agreement and reliability of knower-level groups[5] (non-knowers, subset-knowers, CP-knowers). Here, we found an agreement of 86%, and a weighted Kappa of 0.844, which is considered excellent. This suggests that children classified as subset-knowers in the first assessment are likely to remain subset-knowers in the second assessment (and so are non-knowers and CP-knowers).

Overall, these results using the kappa statistic are conceptually identical to those obtained in Exp1. Specifically, we found that the reliability of Give-N non-titrated, when considering all knower-levels at once, is high, but that this effect is likely driven by the high reliability of non-knowers and CP.

Next, we assessed whether there was an order effect, whenever knower-levels did not match across the two tasks.

[5] In task 1, there were 6 children classified as non-knowers, 47 subset-knowers (1K to 4K) and 28 CP-knowers. In task 2, there were 5 non-knowers, 46 subset-knowers and 30 CP-knowers.

As in Experiment 1, more children decrease their knower-level from Give-N 1 to Give-N 2 (decreased n=13; increase n=9) but this difference was not significant (p=0.40). Also, as in Experiment 1, most of these children had knower-levels that differed by one level (difference of 1 level, n=16; difference of 2 levels, n=4; difference of 3 levels, n=2; p=0.26).

Table 2: Distribution of Knower-Levels at the first (T1) and second (T2) assessment of Give-N non-titrated

	0K	1K	2K	3K	4K	CP
T1	6	18	10	7	12	28
T2	5	21	11	10	4	30

Testing Location We found no difference in agreement between knower-levels depending on the testing location (in lab vs offsite; p=0.57)

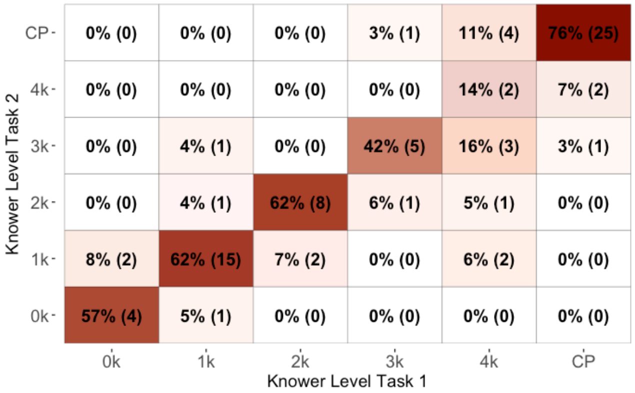

Figure 2: Knower-level classification in the first (T1; x axis) and second assessment (T2; y axis) of Give-N non-titrated. The percentages represent the percent effective agreement – i.e., the agreement calculated over not all paired knower-levels, but those paired knower-levels in which at least one belongs to the knower-level in consideration. The number in parenthesis represent the frequency of the paired knower-level. The color scale is based on the proportion of effective agreement, where darker red represent higher agreement.

Comparing the reliability of the two Give-N types To investigate the difference in rates of agreement across Give-N type (titrated and non-titrated) and knower-levels, we ran a logistic model using glm function in R (R Core Team, 2017). In a first model, we predicted agreement (coded as yes or no) from Age (in months) and Highest Count[6], but both factors were not significant (both ps>0.05). Because these factors were not significant, we did not add them into our principal model of interest. In our principal model, we predicted agreement from Give-N type – either titrated or non-titrated – knower-level group (i.e., subset-knower, non-knower and CP) and the interaction between the two factors[7]. In this model, only the main effect of knower-level group was significant, when considering the knower-level values of both Give-N T1 and T2 (both ps<0.05). In other words, a child that was classified as a subset-knower, either when considering T1 or T2, was less likely to have an agreement between Give-N assessments compared to children classified as non-knowers or CP-knowers in at least one of the two assessments. This result corroborates the findings of Experiments 1 and 2 using the kappa index.

General Discussion

The goal of this study was to investigate the reliability of both versions of Give-N, titrated (Experiment 1) and non-titrated (Experiment 2). In both experiments, when considering all knower-levels together, we found an overall high reliability of the Give-N task. This suggests that children who were classified into a particular N-knower-level in their first assessment were likely to receive the same knower-level assignment in the second assessment. However, we also found evidence that the reliability of the task was somewhat affected by individual knower-level group. Specifically, for both the titrated and non-titrated type, the reliability of the subset-knower group was lower than that of non-knowers and CP-knowers, suggesting that the high reliability of the task might be driven by these last two knower-level groups. In line with this result, we also found that knower-level group (either subset-knower, non-knower or CP) was a significant predictor of agreement, regardless of the Give-N methodology used. Finally, the testing location (either in lab or off-site) didn't have any impact on the rates of agreement of the task, for either the titrated or non-titrated version.

Overall, these results bring encouraging news to researchers using Give-N to study number words comprehension in children, as we show that Give-N has a high and satisfactory reliability. Nonetheless, these findings have practical implications for how future studies should use this task. Given the lower reliability of individual knower-levels within the subset knower group, researchers could try to use a broader knower-level group distinction (e.g., 3 groups: non-knowers, subset-knowers and CP), as an alternative to individual knower-levels to predict outcomes. The reliability of these groups (e.g., that a child classified as a subset-knower at T1 remains a subset-knower at T2) was in fact the highest obtained for both Experiments 1 and 2. In cases where using the broader distinction is not applicable however, for example in studies investigating questions that are specifically about individual knower-levels (e.g., Almoammer et al., 2013), researchers could use the reliability index provided in this study in order to estimate the sample size needed to reach adequate power. In addition, since there was no difference in agreement and apparent

[6] On average, children's counting skills were highly variable (*M* = 12.8; *SD* = 13.0; range = 0-100).

[7] The first model specification was Agreement ~ Highest Count + Age. The second model was Agreement ~ Give-N type * KL group.

reliability across the titrated and non-titrated versions of Give-N, it might be more advantageous for researchers to use the titrated version of Give-N in a study as this version is faster to run than the non-titrated one (~8min for titrated vs ~10 minutes for non-titrated), and therefore, more appropriate for younger children with a limited attention span.

An interesting theoretical question raised by these findings is why there is variation in reliability across individual knower levels and what can this tell us about models of number words acquisition? We address this question by looking at non-knowers, subset-knowers and CP individually. For non-knowers, our results show that these children's behaviors are consistent over repeated assessment; non-knowers tend to grab all the objects or provide quantities somewhat randomly. This is compatible with a view in which these children don't have yet a reliable hypothesis for the meaning of any number words. On the opposite end of the spectrum, children who understand the meaning of counting (at least to 6; CP-knowers) are also consistent in how they perform at Give-N and can accurately count objects as they provide them. Subset-knowers, on the other hand, are less consistent in their responses to requests, as demonstrated by the lower reliability across Give-N tasks. The interesting puzzle is why there is a high variability in the reliability measures within subset-knowers and what this variability can tell us. One possible explanation is that subset-knowers learn number words gradually, and have solid knowledge of some numbers, but only partial, instable, knowledge of larger ones. Evidence for this comes from the fact that subset-knowers perform slightly better than chance when asked to give numbers just beyond their knower level (Barner & Bachrach, 2010; Gunderson, Spaepen, & Levine, 2015; Wagner, Chu, & Barner, 2019). The possibility that an N-knower might have partial knowledge of N+1 or N+2 might explain why these children can be classified, just by chance, as N-knower at T1 and then N+1-knower at T2. Such an explanation would provide support for models of number words acquisition on which children begin learning all small numbers simultaneously, but acquire adult-like meanings at different moments due to differences in frequency. Children may learn the meaning of "one" earlier than "two" not because it is easier or required for learning larger numbers, but simply because they hear it much more frequently (Dehaene & Mehler, 1992).

Given the variation in reliability for subset stages, another interesting question raised by these results is how necessary it is for children to have fixed association between number words and the non-symbolic representations of those number words in order to learn the meaning of counting. Our data support the view that there is an important conceptual distinction between subset-knowers and CP-knowers, but doesn't specify what explains this difference. Given the high variability in responses within subset-knowers it is possible that this variability remains the same when children become CP-knowers, but that what characterizes CP-knowers is the mastery of a counting procedure. In other words, CP-knowers may be distinct from subset-knowers only in that they can apply a procedure without implicating representations of small number words. Future studies should explore this possibility.

While this study is the first to assess the reliability of Give-N in a systematic manner, our results leave open multiple questions. For example, it would be interesting to not only assess the reliability of the titrated and non-titrated Give-N tasks in a within-subject design, but also to investigate the validity of different Give-N versions by relating them to other tasks frequently used in the numeracy literature. Our lab is currently addressing this question, by testing children with both the titrated and non-titrated versions of Give-N, as well as the What's On this Card task adapted from Le Corre & Carey (2007).

Finally, these results can have implications not only for the development of numeracy but also for other aspects of cognitive development. Specifically, given the increasing concern about replicability in research, it is important to assess and discuss measurements and methodological differences across labs. As noted in the introduction, there are many differences in the way that tasks can be administered across labs and experimenters, including the type of questions asked, the order of trials, the type of material and location. Here, we provide evidence that testing location doesn't impact the reliability of Give-N, which is good news for researchers using this task. Addressing the potential impact of these factors might be a step to better understand some of the issues at the core of the replicability crisis.

Acknowledgments

We would like to thank members of the Language and Development Lab for their support during this project and the reviewers for their helpful comments and feedback.

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*, 201313652.

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive psychology, 60*(1), 40-62.

Chu, F. W., vanMarle, K., & Geary, D. C. (2016). Predicting children's reading and mathematics achievement from early quantitative knowledge and domain-general cognitive abilities. *Frontiers in psychology*, *7*, 775.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition, 43*(1), 1-29.

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

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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

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., Li, P., Huang, B. H., Jia, G., & Carey, S. (2016). Numerical morphology supports early number word learning: Evidence from a comparison of young Mandarin and English learners. *Cognitive psychology*, *88*, 162-186.

Negen, J., & Sarnecka, B. W. (2012). Number-concept acquisition and general vocabulary development. *Child development*, *83*(6), 2019-2027.

Nunnally Jr, J. C. (1970). Introduction to psychological measurement.

Piantadosi, S. T., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental Science*, *17*(4), 553-563.

Purpura, D. J., & Simms, V. (2018). Approximate number system development in preschool: What factors predict change?. *Cognitive Development*, *45*, 31-39.

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

Sarnecka, B. W., Kamenskaya, V. G., Yamana, Y., Ogura, T., & Yudovina, Y. B. (2007). From grammatical number to exact numbers: Early meanings of 'one','two', and 'three'in English, Russian, and Japanese. *Cognitive psychology*, *55*(2), 136-168.

Schaeffer, B., Eggleston, V. H., & Scott, J. L. (1974). Number development in young children. *Cognitive Psychology*, *6*(3), 357-379.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, *85*(3), 257-268.

Team, R. C. (2017). R Core Team (2017). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria. URL http://www. R-project. org/., page R Foundation for Statistical Computing*.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on psychological science*, *4*(3), 274-290.

Wagner, K., Chu, J., & Barner, D. (2019). Do children's number words begin noisy?. *Developmental science, 22*(1), e12752.

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

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.