Cognition 222 (2022) 104998

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Assessing the knower-level framework: How reliable is the Give-a-Number task?

Elisabeth Marchand *, Jarrett T. Lovelett, Kelly Kendro, David Barner

[a] Department of Psychology, University of California San Diego, United States of America

ARTICLE INFO

Keywords:
Give-a-Number task
Reliability
Validity
Number acquisition
Assessment
What's-On-This-Card

ABSTRACT

The Give-a-Number task has become a gold standard of children's number word comprehension in developmental psychology. Recently, researchers have begun to use the task as a predictor of other developmental milestones. This raises the question of how reliable the task is, since test-retest reliability of any measure places an upper bound on the size of reliable correlations that can be found between it and other measures. In Experiment 1, we presented 81 2- to 5-year-old children with Wynn (1992) titrated version of the Give-a-Number task twice within a single session. We found that the reliability of this version of the task was high overall, but varied importantly across different assigned knower levels, and was very low for some knower levels. In Experiment 2, we assessed the test-retest reliability of the non-titrated version of the Give-a-Number task with another group of 81 children and found a similar pattern of results. Finally, in Experiment 3, we asked whether the two versions of Give-a-Number generated different knower levels within-subjects, by testing 75 children with both tasks. Also, we asked how both tasks relate to another commonly used test of number knowledge, the "What's-On-This-Card" task. We found that overall, the titrated and non-titrated versions of Give-a-Number yielded similar knower levels, though the non-titrated version was slightly more conservative than the titrated version, which produced modestly higher knower levels. Neither was more closely related to "What's-On-This-Card" than the other. We discuss the theoretical and practical implications of these results.

1. Introduction

Over the past 40 years, a large corpus of studies has shown that children acquire the meanings of number words in a predictable and protracted stage-like sequence. This developmental sequence has been revealed in large part by a single measure of number word knowledge, called the Give-a-Number task (Give-N). Though versions of this task were used as early as the 1970s to study number word comprehension (Schaeffer, Eggleston, & Scott, 1974), Give-N emerged as a type of gold standard after it was used by Wynn (1990, 1992) to describe children's progression through stage-like "knower levels" in both cross-sectional and longitudinal designs. In the task, an experimenter provides children with a set of small counters (e.g., 10–15 toy apples), and asks them to give specific numbers of things, often starting with 1 – e.g., "Can you put *one* apple in the plate?". Children who can consistently give 1 when asked for *one*, but who give inconsistent amounts of objects for other requests are typically called 1-knowers. Similarly, 2-knowers can give 1 and 2 when asked for these quantities but are unable to consistently give appropriate quantities for larger numbers like *three, four,* etc. Following a similar pattern, children go through the stages of 3-knower and sometimes 4-knowers, too. Sometime between the ages of 3;6 and 5, children appear to make a breakthrough, and begin to use counting to correctly give larger sets, at which point they are called "Cardinal Principle knowers" or CP-knowers. This basic developmental pattern appears to be highly replicable across multiple labs in different countries (Almoammer et al., 2013; Barner, Chow & Yang, 2009; Ceylan & Aslan, 2018; Condry & Spelke, 2008; Davidson, Eng & Barner, 2012; Jara-Ettinger, Piantadosi, Spelke, Levy & Gibson, 2017; Le Corre & Carey, 2007; Le Corre, Li, Huang, Jia & Carey, 2016; Le Corre, Van de Walle, Brannon & Carey, 2006; Li, Le Corre, Shui, Jia & Carey, 2003; Nikoloska, 2009: Marchand & Barner, 2019; Meyer, Barbiers & Weerman, 2020; Negen & Sarnecka, 2012; Piantadosi, Jara-Ettinger & Gibson, 2014; Sarnecka & Carey, 2008; Sarnecka, Kamenskaya, Yamana, Ogura & Yudovina, 2007; Sarnecka & Lee, 2009; Sarnecka, Negen & Goldman,

* Corresponding author at: Department of Psychology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92161, United States of America.
E-mail address: emarchan@ucsd.edu (E. Marchand).

https://doi.org/10.1016/j.cognition.2021.104998
Received 2 July 2021; Received in revised form 21 November 2021; Accepted 22 December 2021
Available online 7 February 2022
0010-0277/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2018; Schneider et al., 2020; Wynn, 1990, 1992; Wagner, Kimura, Cheung & Barner, 2015; Spaepen, Gunderson, Gibson, Goldin-Meadow & Levine, 2018; Slusser, Ditta & Sarnecka, 2013). This is important not only because of the theoretical implications of the observed stages (e.g., Carey & Barner, 2019; Le Corre et al., 2006; Le Corre & Carey, 2007; Piantadosi, Tenenbaum & Goodman, 2012; Sella, Slusser, Odic & Krajcsi, 2021), but also because the stages provide a framework for comparing data across studies and across cultures. Numerous studies have now tested how vocabulary size, grammatical cues, and other cultural factors relate to different knower level stages (Almoammer et al., 2013; Barner, Libenson, Cheung & Takasaki, 2009; Le Corre et al., 2016; Marušič et al., 2016; Negen & Sarnecka, 2012; Sarnecka et al., 2007, 2018), and others have asked how knower levels relate to later mathematics achievement (Chu, vanMarle & Geary, 2016; Geary & Vanmarle, 2016; Moore, VanMarle & Geary, 2016; Purpura & Simms, 2018; Spaepen et al., 2018) or the development of other cognitive processes (Abreu-Mendoza, Soto-Alba & Arias-Trejo, 2013; Le Corre, 2014; Mussolin, Nys, Content & Leybaert, 2014; Sarnecka & Wright, 2013; Shusterman, Slusser, Halberda & Odic, 2016).

Critically, however, the replicability of the overall knower level framework does not itself assure the reliability of individual knower level classifications and doesn't guarantee that testing correlations between knower levels and other factors will generate meaningful results. Currently, the reliability of the Give-N task is not known. This is important because the strength of a correlation between two observations (e.g., knower level and vocabulary size), *r*(ObservedA,ObservedB), is bounded not only by the true correlation between the true value of the variables being measured, *r*(TrueA,TrueB), but also by the test-retest reliability of these measures taken individually, reliabilityA, reliabilityB (Nunnally, 1970).

$$r(\text{ObservedA}, \text{ObservedB}) = r(\text{TrueA}, \text{TrueB}) \text{ x} \sqrt{(\text{reliabilityA x reliabilityB})}$$

Thus, as noted by Vul, Harris, Winkielman, & Pashler (2009), in a scenario in which a true correlation between two variables is 100% but the test-retest reliability is 0.7 for one measure and 0.8 for the second, the highest detectable correlation should be 0.75 (i.e., 1 x $\sqrt{(0.7 \times 0.8)}$). In the current context, this means that if individual knower levels (e.g., the 1-knower stage) exhibit very low reliability (e.g., 0.3), then the size of expected correlations between this knower level and other variables should also be low. Consequently, very low reliability would draw into question the validity of knower levels, since the validity of a measure is defined by its ability to make predictions about the outcomes of other measures.[1] More generally, the interpretation of knower level assignments as correlates of 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 three experiments. In Experiment 1, we assessed the test-retest reliability of Wynn's titrated version of Give-N. In the titrated Give-N task, trials are structured such that if a child responds correctly to a request (e.g., giving exactly 1 object when asked for *one*), they are then tested with the next largest number (e.g., *two*), whereas if they fail, they are tested on a smaller number (or again on *one*). This procedure is then repeated until the experimenter can identify the largest number known by the child. In Experiment 2, we investigated the test-retest reliability of an alternative version of Give-N that uses a non-titrated trial structure 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, using the same criteria to identify children's knower levels. We expected that this version might offer stronger reliability than the titrated version, because it features more trials and uses the same trial structure on each testing occasion, unlike the titrated version.[2] In both Experiments 1 and 2, we also considered the role that testing environment might play in the reliability of Give-N by evaluating children in two different settings – either in the lab or outside of the lab (e.g., in a museum, preschool, etc.) – as some studies have reported different outcomes in these different settings (Newman, Dickstein & Gargan, 1978; Rasmussen, Keene, Berke, Densley & Loof, 2017; Yantz & McCaffrey, 2009; cf. Pfefferle, Machen, Fields & Posnick, 1982). Finally, in Experiment 3, we compared the titrated and non-titrated Give-N versions within-subjects, to determine whether they generated different results, and whether either of the two was more conservative (e.g., by ascribing less knowledge). Also, Experiment 3 attempted to probe how the two Give-N methods are related to another frequently used measure of number knowledge by comparing them to the What's-on-this-Card task, which assesses how accurately children label sets when presented visually.

2. Experiment 1: Give-a-Number titrated

2.1. Method

2.1.1. Participants

We tested 106 English-speaking children. A total of 25 children were excluded from analysis because of (1) failure to complete all 3 tasks ($n = 11$), (2) language delay ($n = 1$), (3) being a non-English primary speaker ($n = 2$), (4) falling outside the targeted age range ($n = 4$) or (5) experimenter error ($n = 7$). Our final sample included 81 children, aged 2;2 to 4;1-year-old ($M = 3;3$ years). We chose to test participants in this age range as previous studies suggest that it features the most variability in knower levels. Participants were recruited from a parent database (lab), preschools, and museums in San Diego, California, spanning a wide range of socioeconomic backgrounds. Informed consent was obtained from parents. The study received approval by the institutional ethics committee of UCSD.

2.1.2. Materials and procedure

Children were tested either in the lab or offsite at museums and preschools. The testing environment in museums and preschools was similar and consisted of a relatively quiet corner of a room made available by staff. The testing environment in the lab was more quiet than off-site and possible distractions were limited. Each session lasted approximately 8 min and included three tasks administered in the same order for all participants: (1) Give-a-Number task 1, (2) Highest Count task and (3) Give-a-Number task 2. Children received a small prize for their participation at the end of the testing session.

2.1.2.1. Titrated Give-a-Number task. This task was based on Wynn (1992). Stimuli included a puppet, a plastic plate, and a pile of small plastic toys. Participants 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 on the plate for Mr. Monkey, ok? Listen carefully! Can you put N banana(s) on the plate?* (N is the number word). *Put N banana(s) on the plate and tell me when you're all done.*" Following these instructions, children were asked to count to verify that

[1] As explained in Buelow (2020): "A task that is not reliable can not be valid, and lowered reliability can limit inferences made from the task to real-world behaviors." The logic is that an outcome X can't predict a second outcome Y if it can't predict itself (i.e., if it is unreliable). And if X can not explain properties of the world, then it is not a valid measure.

[2] We reasoned that additional trials might increase reliability by providing more information and reducing the likelihood of underestimating (or overestimating) knowledge, and that stable trial structure should reduce the possibility that low reliability is due to variability introduced by differences in methods across testing sessions. Specifically, whereas random performance errors made by children will have no impact on the trial structure of the non-titrated version of the task since its trial structure is predetermined, such errors may significantly change the trial structure of the titrated version, since an error on a trial forces a retreat to a smaller number.

they had provided N (i.e., *"Is that N? Can you count and make sure?"*). If they chose to change their answers, only their final responses were recorded. Participants were always asked for *one* first, and then *two*. If the child succeeded on both trials, the experimenter then asked for *three*. Otherwise, they asked for *one*. The subsequent requests depended on the child's pattern of response: if the child succeeded in providing N items, the experimenter asked for N + 1 and if the child failed, they asked for N - 1. The lowest request was *one* and the highest was *six*. Children were credited as N-knowers (e.g., 2-knowers) if they correctly gave N objects at least 67% of the time when asked for N. Furthermore, to be credited as N-knowers, children needed to use N 67% of the time only for requests of N and not for other requests (in practice, this meant that children could give N only once for requests other than N). Children were credited as CP-knowers if they were able to provide all sets up to *six* based on these criteria, or if they responded to each request (*one* to *six*) consecutively without error, in accordance with Sarnecka & Wright (2013). Aside from this last instance (of CP-knowers), participants were tested with a minimum of 2 trials for *N,* and for numbers tested twice, children needed to succeed on both trials to be tested on the next trial or be credited as N-knower. Children who correctly gave 1 object when asked for *one* (but failed for *two* and larger requests) were classified as 1-knowers. Children who answered successfully for *one* and *two* were credited as 2-knowers and so forth. Although past studies have often classified children who succeed at *five* as CP-knowers, we chose to categorize children as 5-knowers if they succeeded at *five* but failed at *six*. Although more conservative, this criterion allowed us to test the claim that knower levels higher than 4 exist and can be diagnosed (Krajcsi, Fintor & Hodossy, 2018). However, allowing for an additional knower level in the classification risks decreasing the reliability of the task and of some knower levels in particular. Nevertheless, as we report below, including 5-knowers didn't impact the reliability of the task because the number of 5-knowers was low (3 children at T1 and 4 at T2).[3]

2.1.2.2. Highest count (HC). This task was used to verify that our sample was representative of previously reported samples of the same age and served as a filler task between the two 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.

2.1.3. Analyses

The choice of a reliability index depends crucially on the scale of the outcome measure of interest. Cohen's Kappa is very commonly used for nominal scales, especially when the outcome of interest is binary, such as the presence or absence of some clinical condition (Hallgren, 2012). However, the basic computation of Kappa can be modified to weight different disagreements in classification differently, allowing the approach to work for ordinal scales as well (in which, say, the difference between 4 and 2 is larger than that between 4 and 3). Intra-class correlations (ICC; Hallgren, 2012) are designed for use with tasks that produce continuous outcome measures, but also produce interpretable results for ordinal scales. Thus to select a measure of reliability for knower level classifications, one must first decide how to conceptualize that construct: as a smooth continuum of knower levels, or as a discrete set of stages? Is the transition from zero-knower to one-knower a similar jump in number knowledge as the transition from two-knower to three-knower? Because it is not clear whether any single choice of reliability metric is entirely free of drawbacks with respect to complexity of the knower level scale, we introduce and report several different metrics, so that readers may use their own judgment in assessing the degrees of reliability reported here.

Table 1
Example of a simplified contingency table used in the reliability computations.

		Assessment 1	
		1-knower	2-knower
Assessment 2	1-knower	*a*	*b*
	2-knower	*c*	*d*

Note. Example of contingency table with Give-N's One-Knower (1-knower) and Two-knower (2-knower) only.

Here we describe the different reliability indexes used throughout Experiments 1, 2 and 3, including Kappa (weighted and unweighted), Agreement, Bias Index, Prevalence Index, Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) and Intra-class correlation (ICC). The Kappa statistic was preferred for our reliability assessment as it is considered a standardized index of reliability for categorical variables (Hallgren, 2012), with which the knower level framework is more compatible (relative to continuous scales). Across the different measures of reliability for categorical data, we also prioritized the Kappa statistic because Kappa, in its weighted version, is compatible with both nominal and ordinal data, which we used in our experiments, allowing us therefore to provide a consistent measure across different analyses. However, in addition to the weighted Kappa, we provide the reader with Intra-class Correlation when dealing with ordinal data as it is a common measure used in this context and may be preferred by researchers who conceptualize knower levels as a continuous scale (although we do not endorse this practice). All analyses were computed in R (R Core Team, 2018) and Kappa analyses were performed using the "vcd" (Meyer, Zeileis & Hornik, 2021) and epiR packages (Stevenson & Sergeant, 2021). In our main analyses, reliability was measured using the weighted version of the Kappa statistic (Cohen, 1960, 1968), defined in the following way:

$$K = \frac{Po - Pe}{1 - Pe}$$

In this expression, K represents the Kappa statistic, Po is the observed (overall) agreement and Pe the agreement expected by chance. Overall agreement corresponds to the total number of matches between the first and second assessment of a task (i.e., the sum of the values on the diagonal of a contingency table) divided by the total number of observations (see Table 1 for an example of a simplified contingency table). Agreement expected by chance refers to the sum of the theoretical frequencies in each cell of the diagonal, which are calculated using the same formula as for computing expected frequencies for Pearson's Chi square (i.e. by taking the product of observed marginal proportions classified as each knower level across tasks).

In the modified weighted Kappa formula, Po and Pe are calculated using a matrix of (dis-)agreement weights, which specify the degree to which each possible pair of classifications from the two tasks (dis-)agree. In the case of knower levels, this means that the difference between, for example, a 1-knower and a 5-knower can be represented as larger than the difference between a 1-knower and a 2-knower. That feature enables weighted Kappa to handle ordinal scales, since it can attach greater weight to large differences between levels than to small differences (Cohen, 1968). Importantly, it is incumbent on the investigator to decide *how much* weight to assign each possible (dis-)agreement, by carefully designing a weight matrix.

In principle, a fully custom weight system could be used to describe the severity of disagreement for each pairwise combination of classifications across the two tasks. For example, disagreements in which a subject is classified once as a CP-knower and once as a non-knower (CP-0k disagreements) could be weighted as arbitrarily more severe than

[3] We re-ran all analyses with 5-knowers categorized as CP-knowers and obtained virtually the same results as presented below; while the reliability for the task overall remained unchanged (linear weight = 0.88 vs 0.87), there was, unsurprisingly, a slight increase of reliability for the CP-knowers (from 0.827 to 0.852), the subset-knowers (from 0.681 to 0.700) and the knower-level groups (0.824 to 0.850) analyses. However, in all of these cases, the increase was negligible.

disagreements in which the particular value of subset-knower was different across tasks. Each other combination of disagreements, such as CP-3K, 0K–4K, 4K–0K, etc., would have to be specified individually. *Any* choice of weighting, especially a custom weight scheme, therefore reflects a judgment regarding the nature of the number word acquisition process and its stages. For that reason, we refrain from developing our own customized weight system (with which other researchers could reasonably disagree). Instead, we report results using two common weighting systems: linear weights (in which the penalty for a disagreement is proportional to the absolute value of the difference in ranks across the two levels), and quadratic weights (in which the penalty is proportional to the square of that difference); using linear weights, a 4K–2K disagreement is twice as severe as a 4K–3K disagreement, while under quadratic weighting, 4K–2K is four times as severe as 4K–3K. Our preference is for linear weights, as that approach makes fewer theoretical assumptions about the trajectory of number knowledge development.

In addition to Kappa, in all analyses we reported either the overall agreement or the effective agreement depending on the data under study. Effective agreement is defined as the number of matches divided by the number of observations that include at least one of the knower levels in consideration. Both overall agreement, the total number of matches over total values, and effective agreement are inflated indexes of reliability because they don't consider the agreement that could have occurred by chance (Luck et al., 2012; Viera & Garrett, 2005), which Kappa (weighted and unweighted) accounts for, making Kappa more conservative than raw measures of agreement.

Some authors have argued that the magnitude of Kappa can be influenced by factors such as prevalence and bias in the data and that consequently, Kappa can be misleading in cases where these factors are considerable (Byrt, Bishop & Carlin, 1993; Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990). Prevalence refers to the relative difference of agreement between raters or tasks across conditions. The prevalence index is calculated in the following way (refer to Table 1):

$$Prevalence\ index = \frac{|a-d|}{n}$$

Where $|a\text{-}d|$ is the absolute value of the difference between the frequencies of cells on the diagonal (agreements; *a* and *d* in the Table 1) and n is the total number of observations.

If the prevalence index is high, suggesting that there is a high asymmetry in the frequencies in the cells of the diagonal, then Kappa will be reduced. The bias effect on Kappa occurs when there is large asymmetry in the frequencies of cells outside the diagonal, in other words, of disagreements (*b* and *c*). A high bias index can lead to an oversized Kappa. The bias index is measured in the following way:

$$Bias\ index = \frac{|b-c|}{n}$$

In our assessment of reliability, alongside Kappa, we provide the prevalence and bias indexes,[4] as well as the Prevalence-adjusted bias-adjusted Kappa (PABAK) coefficient whenever the data allow it. It is important to note, however, that the prevalence and bias indexes are not measures of reliability per se, but instead provide an indication of potentially unbalanced data, and consequently, whether to rely more on PABAK than Kappa when interpreting the results.[5] PABAK is an adjustment of Kappa that takes into account the influence of bias and prevalence. It is calculated by substituting the actual frequencies of cells *a* and *d* by their average to account for prevalence, and by replacing the actual frequencies of cells *c* and *b* by their average to account for bias. Not all studies agree on which Kappa coefficient, the original or the PABAK, should be used as the main reference value. Some argue that bias and prevalence are the inevitable result of the natural disparities in the population under study and that correction coefficients such as PABAK can therefore be misleading. We follow the recommendations of Byrt et al. (1993) and provide the reader with both values (non-adjusted and PABAK) as well as the prevalence and bias indexes, whenever the data allowed, so that the reader can assess reliability based on a holistic evaluation of these measures. Finally, for some analyses when it was applicable, we also provided the ICC which is another commonly-used statistic for ordinal variables (Hallgren, 2012), based on correlations. Our complete datasets are also available in the following repository: **https://osf.io/48mke/.**

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

Knower Levels	0K	1K	2K	3K	4K	5K	CP
Assessment	Number of Participants						
Time 1	9	14	16	5	7	3	27
Time 2	9	15	15	8	6	4	24

Note: 0K refers to non-knower, 1K to 1-knower, 2K to 2-knower, 3K to 3-knower, etc., and CP to Cardinal Principle knower. 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.

2.2. Results

Table 2 shows the distribution of knower levels in the first and second assessment of the titrated Give-N task. On average children could count just above 10 in the Highest Count task ($M = 12.8$), and their counting skills were variable ($range = 0$ to 100; $SD = 13.0$). Seventy-four out of 81 (91%) participants were found to have a highest count greater than their knower level across the two Give-N assessments.

In our first analysis, we included all knower levels (non-knower to CP-knower) in a 7 × 7 contingency table (see Fig. 1) and obtained an overall agreement of 77% and a weighted Kappa ($K_{w\text{-}linear}$) of 0.87 and 0.95 ($K_{w\text{-}quadratic}$; $Kappa_{unweighted} = 0.71$; $Prevalence\ index_{(mean)} = 0.11$, $range = 0\text{–}0.25$; $Bias\ index = 0.09$; $PABAK_{weighted} = 0.73$; $ICC = 0.97$). All statistics are summarized in Table 3. Some researchers attempt to classify reliability scores according to a scale as described in Table 4; according to this scheme, this level of reliability is considered almost perfect (Fleiss, Levin & Paik, 2003; Landis & Koch, 1977). However, because there is disagreement regarding these labels and their utility (e.g., Sim & Wright, 2005), and because we are mainly interested in quantitative impacts of reliability on the size of correlations between measures (rather than qualitative endorsement of particular tasks), we sidestep the significance of these labels in our discussion. As shown in Fig. 1, the rate of effective agreement (in percentage) across different knower levels was highly variable. Effective agreement was relatively high for non-knowers (80%), CP-knowers (76%), 1-knowers (71%), and

[4] Note that for tables larger than 2 × 2, we calculated the prevalence index by taking the average difference (in absolute value) between all numbers in the diagonal paired together (a-d in Table 1). More precisely, we replaced the |a-d| in prevalence formula by the average difference between all numbers on the diagonal of the contingency table under study. For the bias index, we replaced *b* in the formula by the sum of all numbers above the diagonale and *c* by the sum of all numbers below the diagonale. However, the literature on how to calculate these measures for tables larger than 2 × 2 was very sparse and we could not identify any straightforward way to proceed. Calculating the average prevalence and bias seemed like the more reasonable approach but other researchers might disagree. Results for the prevalence and bias indexes, as well as PABAK (which relies on these 2 measures) for large tables (>2 × 2) should therefore be interpreted with caution.

[5] The criteria to classify a prevalence and bias index as too high are subjective and inconsistent. In our data, we noticed that the prevalence index tended to be particularly high in 2 × 2 tables (e.g., 0.78) and in those cases, we favored the PABAK instead of Kappa for our interpretation of the data.

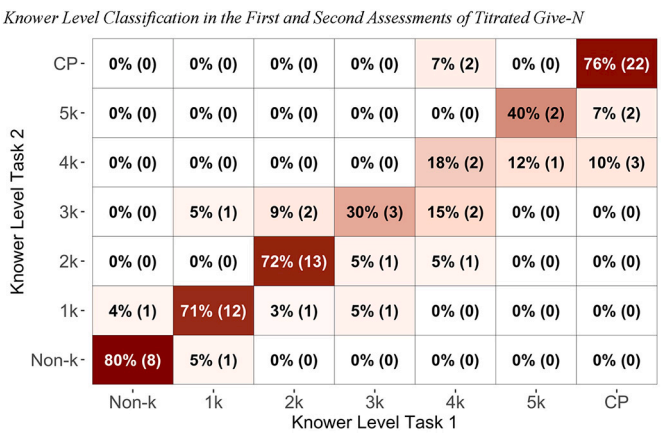

Fig. 1. Knower level classification in the first and second assessments of titrated Give-N.

Note. The first assessment (T1) appears on the x axis, and the second assessment (T2) appears on the y axis. 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 numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement, where darker red represents higher agreement.

2-knowers (72%), but much lower for 3- (30%), 4- (18%) and 5-knowers (40%). Thus, overall, the titrated Give-N task was highly reliable when all knower levels were considered together, although concordance between individual knower levels was lower in some cases.

To further investigate the difference of reliability across individual knower levels, we conducted three follow-up analyses for the subset-knower, non-knower, and CP-knower groups respectively. For the subset-knower analysis, we created a 6 × 6 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-knowers) for Give-N Test 1 (T1) and Give-N Test 2 (T2). We found an effective agreement of 63% and an unweighted Kappa of 0.68 (*Prevalence index*$_{(mean)}$ = 0.15, *range* = 0–0.35; *PABAK* = 0.72).[6] We report the effective agreement here as we were interested in the agreement specifically within the group of subset-knowers and this index does not include non-knowers and CP-knowers. We also report the unweighted Kappa, and not the weighted Kappa, because weighted Kappa assumes an ordered category structure, which is violated by binning non-knowers and CP-knowers into a common category. Next, for the non-knower analysis, we generated a 2 × 2 contingency table with contrasting non-knowers with all other levels for both Give-N T1 and Give-N T2. We obtained an effective agreement of 80% and a Kappa of 0.88 and (*Prevalence index* = 0.78[7]; *Bias index* = 0; *PABAK* = 0.95). Next, for the CP-knower analysis, we created a 2 × 2 table (CP vs non-CP at T1 and T2) and found an effective agreement of 76% and a Kappa of 0.80 (*Prevalence index* = 0.37; *Bias index* = 0.04; *PABAK* = 0.83). These results suggest that the non-knower and CP-knower classifications are highly reliable, and more reliable than classifications within the subset stage (though as already noted, concordance within the subset stage varies between individual levels, as shown in Fig. 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., 1-knower), and more interested in whether they are CP-knowers or subset-knowers. Relatedly, most 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 in T1, was likely to be a subset-knower again in T2. To do this, we divided knower levels into three groups: non-knowers, subset-knowers (1K to 5K) and CP-knowers. We then created a 3 × 3 contingency table with knower level groups at T1 and knower level groups at T2. Here, we found an overall agreement of 89%, and a weighted Kappa (linear) of 0.82 and 0.86 (quadratic; *Kappa*$_{unweighted}$ = 0.80; *Prevalence index*$_{(mean)}$ = 0.28, *range* = 0.17–0.42; *Bias index* = 0.04; *PABAK*$_{weighted}$ = 0.83; *ICC* = 0.92), which is similar to the reliability of all knower levels taken together. This suggests that children who were classified as subset-knowers in the first assessment were very likely to remain subset-knowers in the second assessment, just like non-knowers and CP-knowers.

Next, we asked whether knower levels systematically increased or decreased between T1 and T2. An increase could signal a practice effect while a decrease could suggest a fatigue effect. In total, more children exhibited a decrease in their knower level from T1 to T2 (decreased n = 13; increased n = 6) but this difference was not significant (Wilcoxon rank test; W = 3368.5; p = .76). Furthermore, most of these children had knower levels that differed by one level (difference of 1 level, n = 11; difference of 2, n = 8).

Finally, we assessed whether testing location (either in-lab or off-site) was related to knower level classification. To do so, we conducted an ordinal logistic regression ("porl" function in MASS package in R; Venables & Ripley, 2002) with knower levels as the dependent variable and location as the predictor, which revealed no significant effect of location (t = 0.64; p = .52).[8] We also conducted a Fisher's exact test to see if there was a difference in agreement (i.e., matches vs non-matches) between knower levels at T1 and T2 based on testing location, but this was not the case (p = .43).

2.3. Discussion

In Experiment 1, we found that the titrated Give-N task was highly reliable both when all knower levels were considered at once and when considering knower level groups (i.e., subset-knowers, non-knowers, and CP-knowers). The results using the ICC statistic corroborated these findings. However, we noted substantial variation in the concordance of individual knower levels, particularly within the group of subset-knowers, with relatively high concordance for non-knowers, 1-knowers, 2-knowers, and CP-knowers, but lower concordance for 3-, 4-, and 5-knowers.

3. Experiment 2: Give-a-Number non-titrated

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

3.1. Method

3.1.1. Participants

In total, 101 English-speaking children were tested for this experiment. Twenty children were excluded because of (1) failure to complete all 3 tasks (n = 12), (2) language barrier (n = 1), (3) not being in the targeted age range (n = 5), and (4) experimenter error (n = 2), leaving a final sample of 81 children, aged 2;6 to 4;1-year-old (M = 3;4 years). Children were recruited in the same way as in Experiment 1.

[6] Note here that the Bias Index is not valid in this subset-knowers analysis since the "non-subset-knower" category is not ordered and we would obtain different indexes based on its position in the contingency table, which can be placed arbitrarily either on the right or left of the contingency table.

[7] Note that in this analysis with non-knowers, the prevalence index is notably high and that using the PABAK coefficient as the main measure of reliability is recommended.

[8] This is the result obtained when knower levels at T1 are used as the dependent variable. We obtained the same outcome when using knower levels at T2 (t = −0.06; p = .95).

Table 3
Summary of Reliability measures and coefficients of the Titrated Give-N at T1 and T2 across different knower levels analyses.

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	7 x 7	77%	.87 *(w-l)* *95% CI,* .81 to .93 .95 *(w-q)* *95% CI,* .92 to .98 .71 *(unw)* *95% CI,* .59 to .82	.11 *(M)* *range:* 0 - .25	.09	.73 *(w-l)* *95% CI,* .60 to .85 .73 *(w-q)* *95% CI,* .56 to .89	.97 *95% CI,* .96 to .98
Subset-knower only	6 x 6	63% (effective)	.68 *(unw)* *95% CI,* .56 to .80	.15 *(M)* *range:* 0 - .35	*NA*	.72 *(unw)* *95% CI,* .61 to .83	*NA*
Non-knower vs others	2 x 2	80% (effective)	.88 *(unw)* *95% CI,* .66 to 1	.78	.0	.95 *(unw)* *95% CI,* .83 to .99	*NA*
CP-knower vs others	2 x 2	76% (effective)	.80 *(unw)* *95% CI,* .58 to 1	.37	.04	.83 *(unw)* *95% CI,* .66 to .93	*NA*
Knower-level groups	3 x 3	89%	.82 *(w-l)* *95% CI,* .71 to .94 .86 *(w-q)* *95% CI,* .76 to .95 .80 *(unw)* *95% CI,* .68 to .92	.28 *(M)* *range:* .17 - .42	.04	.83 *(w-l)* *95% CI,* .72 to .94 .83 *(w-q)* *95% CI,* .71 to .96	.92 *95% CI,* .88 to .95

Note: The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

Table 4
Interpretation of Kappa Based on Landis and Koch (1977)'s Scale.

Kappa	Interpretation
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Note. A Kappa measure of 0 equates chance while a Kappa of 1 equates a perfect agreement. Negative Kappa measures are possible and represent less agreement than chance (i.e., disagreement).

3.1.2. Materials and procedure

The testing environments were the same as in Experiment 1, except that children were presented with a non-titrated version of Give-N, twice, separated by the Highest Count task. Because the non-titrated version of Give-N includes a fixed number of trials, each session lasted approximately 10 min, slightly longer than in Experiment 1.

3.1.2.1. Non-titrated Give-a-Number task. This task was identical to the titrated version used in Experiment 1, except for the trial structure. Children were given 15 trials including three for each of the numbers 1, 2, 3, 4, and 6. Note that since we did not ask for *five*, children could not be classified as 5-knowers in this version, unlike in the titrated task (though, in Experiment 1, only 5 children were ever classified as 5-knowers). We created two lists of trials in a pseudorandom order. All children were presented with both lists counterbalanced in order at T1 and T2 across children. The criteria to assign knower levels were the same as those used in the titrated version, with an emphasis on the requirement for children to succeed at all numbers below N to be credited as N-knowers (i.e., children couldn't skip some numbers). Children were credited as CP-knowers if they could correctly give six on two out of three trials.

3.1.2.2. Highest count (HC). The task was identical to Experiment 1.

3.2. Results

Table 5 shows the distribution of knower levels in the first and second assessments of the task. As in Experiment 1, most children counted just above 10 in the Highest Count task ($M = 13.6$), and their counting skills were variable (*range* = 1 to 59; $SD = 12.2$). Seventy-seven children out of 81 (95%) had a highest count higher than their knower level across the two Give-N tasks.

Table 5
Distribution of knower levels at the first and second assessments of non-titrated Give-N.

Knower Levels	0K	1K	2K	3K	4K	CP
Assessment	Number of Participants					
Time 1	6	18	10	7	12	28
Time 2	5	21	10	11	4	30

Note: 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.

Table 6
Summary of Reliability measures and coefficients of the Non-Titrated Give-N at T1 and T2 across different knower levels analyses.

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	72%	.81 *(w-l)* *95% CI,* .73 to .89 .90 *(w-q)* *95% CI,* .84 to .96 .63 *(unw)* *95% CI,* .51 to .75	.12 *(M)* *range:* .03 - .28	.04	.66 (*w-l*) *95% CI,* .52 to .80 .66 (*w-q*) *95% CI,* .48 to .84	.95 *95% CI,* .92 to .97
Subset-knower only	5 x 5	56% (effective)	.61 *(unw)* *95% CI,* .48 to .74	.16 *(M)* *range:* .04 - .33	*NA*	.65 *(unw)* *95% CI,* .52 to .77	*NA*
Non-knower vs others	2 x 2	57% (effective)	.71 *(unw)* *95% CI,* .49 to .92	.86	.01	.93 *(unw)* *95% CI,* .79 to .98	*NA*
CP-knower vs others	2 x 2	76% (effective)	.79 *(unw)* *95% CI,* .57 to 1	.28	.02	.80 *(unw)* *95% CI,* .63 to .91	*NA*
Knower-level groups	3 x 3	86%	.77 *(w-l)* *95% CI,* .64 to .90 .80 *(w-q)* *95% CI,* .69 to .91 .75 *(unw)* *95% CI,* .61 to .89	.31 *(M)* *range:* .20 - .46	.04	.80 *(w-l)* *95% CI,* .68 to .92 .80 *(w-q)* *95% CI,* .66 to .94	.89 *95% CI,* .83 to .93

Note: The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

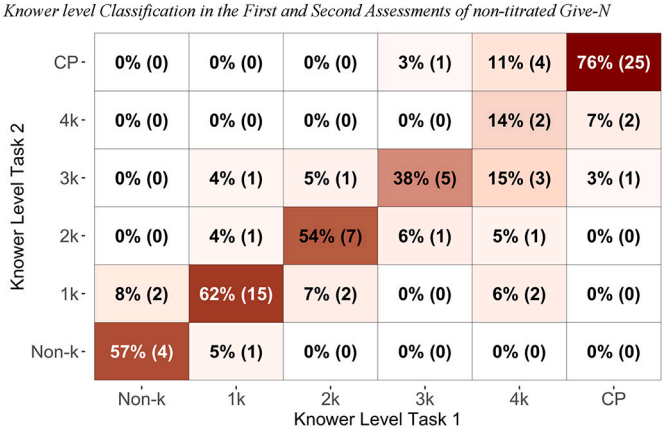

Fig. 2. Knower level Classification in the First and Second Assessments of non-titrated Give-N.
Note. The percentages represent the percent effective agreement – i.e., the agreement calculated over not all paired knower levels, but those paired knower levels in which at least one belongs to the knower level in consideration. The number in parentheses represents the frequency of the paired knower level. The color scale is based on the proportion of effective agreement, where darker red represents higher agreement.

We first calculated the reliability of the non-titrated task, including all knower levels (0 to CP) in a 6 × 6 contingency table.[9] All statistics are summarized in Table 6. We found an overall agreement of 72% and a $K_{w\text{-}linear}$ of 0.81 and 0.90 (quadratic; $Kappa_{unweighted} = 0.63$; *Prevalence index*$_{(mean)} = 0.12$, *range* = 0.03–0.28; *Bias index* = 0.04; $PABAK_{(weighted)} = 0.66$; *ICC* = 0.95). Fig. 2 illustrates the contingency table used and the knower levels at T1 and T2 as well as their effective agreement.

Next, we explored the reliability for subset-knowers, non-knowers and CP-knowers separately. All Kappas were unweighted in these analyses. For the subset-knower analysis, we created a 5 × 5 contingency table with knower levels 1 to 4 and a non-subset-knower category at T1 and T2. We found an effective agreement of 56% and a Kappa of 0.61 (*Prevalence index*$_{(mean)} = 0.16$, *range* = 0.04–0.33; *PABAK* = 0.65).[10] In the non-knower analysis (2 × 2 contingency table), we obtained an effective agreement of 57% and a Kappa of 0.71 (*Prevalence index* = 0.86; *Bias index* = 0.01; *PABAK* = 0.93). In the CP-knower analysis, we found an effective agreement of 76% and a Kappa of 0.79 (*Prevalence index* = 0.28; *Bias index* = 0.02; *PABAK* = 0.80). So far, the results of Experiment 2 are similar to those of Experiment 1; the overall reliability of the task was high but varied across individual knower levels, especially within the group of subset-knowers.

Next, we assessed the agreement and reliability of assignment to

[9] Note that since we did not test for 5, children could not be classified as 5-knower, reducing the knower level categories from 7 (0 to CP; as in Exp 1) to 6.
[10] Note here that, since the "non-subset-knower" category can be placed arbitrarily on either side of the contingency table, the Bias Index is not valid in this analysis.

Table 7
Distribution of Knower Levels for titrated Give-N, non-titrated Give-N and What's-On-This-Card.

Knower Levels	0K	1K	2K	3K	4K	CP
	Number of Participants					
Titrated Give-N	7	20	17	10	5	16
Non-titrated Give-N	10	25	13	8	6	13
What's-On-This-Card	5	23	12	5	11	19

Note: In titrated Give-N, there were 7 children classified as non-knowers, 52 subset-knowers and 16 CP-knowers. In non-titrated Give-N, there were 10 non-knowers, 52 subset-knowers and 13 CP-knowers. In What's-On-This-Card, there were 5 non-knowers, 51 subset-knowers and 19 CP-knowers.

broader knower level groups - i.e., non-knower, subset-knower, or CP-knower. Here, we found an overall agreement of 86%, and a $K_{w\text{-}linear}$ of 0.77 and 0.80 (quadratic; $Kappa_{unweighted} = 0.75$; $Prevalence\ index_{(mean)} = 0.31$, $range = 0.20–0.46$; $Bias\ index = 0.04$; $PABAK = 0.80$; $ICC = 0.89$). This suggests that children classified as subset-knowers in the first assessment were likely to remain subset-knowers in the second assessment (as were non-knowers and CP-knowers). The ICC analysis also confirmed these results.

Next, we assessed whether there was an effect of task order. As in Experiment 1, slightly more children showed a decrease in knower level from T1 to T2 (decreased $n = 13$; increased $n = 10$) but this difference was not significant (Wilcoxon rank test; $W = 3330.5$; $p = .86$). Also, whenever there was a difference of knower levels, more children had knower levels that differed by only one level ($n = 17$) compared to 2 ($n = 4$) or 3 levels ($n = 2$). We also investigated whether there was an effect of trial order across the two tasks (e.g., whether children provided correct responses more frequently for trials presented in the first position vs. trials presented in third position). However, there was no such effect ($z = -1.11$, $p = .27$).

Finally, we found no difference in the distribution of knower levels ($t = -0.064$; $p = .95$) or agreement ($p = .467$) depending on testing location (in-lab vs. off-site).

3.3. Discussion

Against our expectation that the non-titrated Give-N would yield more reliable outcomes, the pattern of results of Experiment 2 is similar to that found in Experiment 1. Specifically, we found that the reliability of the non-titrated Give-N task was high when considering all knower levels at once, but that there was considerable variability when looking at knower levels individually. While the reliability for non-knowers was particularly high, the concordance within the group of subset-knowers varied considerably and was higher for early subset-knowers (1- and 2-knowers) than late subset-knowers (3- and 4-knowers). To better understand how the titrated and non-titrated Give-N versions compare to each other, in Experiment 3 we asked whether they would generate the same knower level within participants. Also, we asked how performance at the two versions of Give-N compared to performance on the *What's-On-This-Card* task (Gelman, 1993; Le Corre et al., 2006).

4. Experiment 3: Give-a-Number titrated, non-titrated and What's-On-This-Card

The results of Experiments 1 and 2 suggest that both versions of Give-N have an overall high test-retest reliability. However, this high reliability does not necessarily mean that the two versions converge on the

Table 8
Reliability measures and coefficients between the Titrated and Non-Titrated Give-N across different knower levels analyses.

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	59%	.70 *(w-l)* *95% CI,* .60 to .80 .84 *(w-q)* *95% CI,* .76 to .92 .49 *(unw)* *95% CI,* .35 to .62	.08 *(M)* *range:* 0 - .16	.17	.50 (*w-l*) *95% CI,* .34 to .67 .50 (*w-q*) *95% CI,* .29 to .72	.91 *95% CI,* .86 to .95
Subset-knower only	5 x 5	47% (effective)	.46 *(unw)* *95% CI,* .31 to .60	.10 *(M)* *range:* .03 - .19	*NA*	.48 *(unw)* *95% CI,* .34 to .62	*NA*
Non-knower vs others	2 x 2	31% (effective)	.41 *(unw)* *95% CI,* .18 to 63	.77	.04	.76 *(unw)* *95% CI,* .57 to .89	*NA*
CP-knower vs others	2 x 2	71% (effective)	.79 *(unw)* *95% CI,* .56 to 1	.61	.04	.87 *(unw)* *95% CI,* .70 to .96	*NA*
Knower-level groups	3 x 3	81%	.64 *(w-l)* *95% CI,* .47 to .81 .69 *(w-q)* *95% CI,* .54 to .84 .60 *(unw)* *95% CI,* .42 to .79	.36 *(M)* *range:* .11 - .55	.08	.72 *(w-l)* *95% CI,* .58 to .86 .72 *(w-q)* *95% CI,* .55 to .89	.82 *95% CI,* .71 to .89

Note: The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

same knower levels when tested within-subjects, since a given task can be reliable despite exhibiting bias. Given this, it is possible that one version generates higher knower levels than the other. For example, because of how knower levels are defined by Wynn's criteria, the inclusion of more trials in the non-titrated version may result in a more conservative - and therefore lower - knower level estimate. Specifically, because a child is only considered an N-knower if they give N for requests of N but not for larger numbers, the inclusion of a greater number of trials creates greater opportunity for random error, possibly resulting in lower knower level estimates. Concretely, if a child correctly gives 3 objects when asked for *three* on both tasks, but then gives 3 objects on ¼ of remaining trials for larger numbers, this will not impact their knower level assignment when using the titrated method (which may include as few as 2 trials above their knower level). However, the knower level may be impacted when using the non-titrated method - e.g., if the child receives 3 trials for each of tested with *four, six,* and *eight,* since ¼ of 9 these trials (i.e., ~2) would constitute 50% of all trials in which 3 is given by the child, ruling out the classification of the child as a 3-knower according to the criteria described above.

While differences between the two Give-N versions can be assessed by directly comparing their outputs, another approach is to ask how each task relates to independent measures of number knowledge. Although there is no single task that tests exactly the same construct as Give-N, a closely related measure is the What's-On-This-Card task (Gelman, 1993; Le Corre et al., 2006), in which children are presented with cards depicting images of sets and are asked to report how many objects they see. In Experiment 3 we administered the What's-On-This-Card task and paired it with a within-subjects comparison of performance on the titrated and non-titrated versions of Give-N. This allowed us to test whether either version of the Give-N task was more closely related to an independent test of number word knowledge.

4.1. Method

4.1.1. Participants

In total, 96 English-speaking children were tested in this experiment. Twenty-one children were excluded from analysis because of (1) failure to complete all 4 tasks ($n = 13$), (2) not being a native speaker of English ($n = 1$), (3) not being in the targeted age range ($n = 1$), (4) experimenter error ($n = 3$) and (5) classifying as a 5-knower in the titrated Give-N ($n = 3$), leaving a final sample of 75 children, aged 2;1 to 4;0-year-old ($M = 3;2$ years). Because there was no difference between testing sites in Experiments 1 and 2, all children in this experiment were recruited offsite (i.e., preschools and museums).

4.1.2. Materials and procedure

Each session lasted approximately 15 min and included (1) Give-a-Number task 1, (2) Highest Count task, (3) Give-a-Number task 2, and (4) What's-On-This-Card task. All participants were administered the tasks in this order, but the order of the titrated and non-titrated Give-N tasks was counterbalanced across children. The procedures for the titrated and non-titrated tasks were identical to what is reported in Experiments 1 and 2, as were the procedures for the Highest Count task.

4.1.2.1. What's-On-This-Card (WOC). This task was modeled after Le Corre et al. (2006). Children were presented with 15 cards containing either 1, 2, 3, 4, or 6 items (balloon, car, dog), assessed 3 times each. Children were asked to report how many items they saw on each card in the following way: "*Now, I'm going to show you some pictures. Your job is to tell me what you see in these pictures. How many [item(s)] do you see in this picture?*". After this initial question, if children did not spontaneously count the items 2 to 6, they were prompted to do so ("*Can you count them for me?*"). Items were aligned and displayed in either one or two rows (depending on the number) and varied in color to maintain children's interest. Two lists of trials in a pseudorandom order were randomly assigned to participants. Before the 15 critical trials, children were presented with a practice trial which was intended to model the expected response and encourage children to provide a number word. We used the same criteria as in Experiments 1 and 2 to assign knower levels; namely that children needed to provide correctly N two out of three times when asked for N, and do so only for N and numbers below. Children were credited as CP-knowers if they could correctly give six on two out of three trials.

4.2. Results

We first assessed the relatedness of the titrated and non-titrated Give-N measures. Although comparing performance on these two versions of Give-N does not strictly amount to assessing reliability - since they are not the same measure - we nevertheless used the statistical tools introduced in Experiments 1 and 2 since measures of reliability offer the best way to assess how often two tasks exhibit agreement in this context.

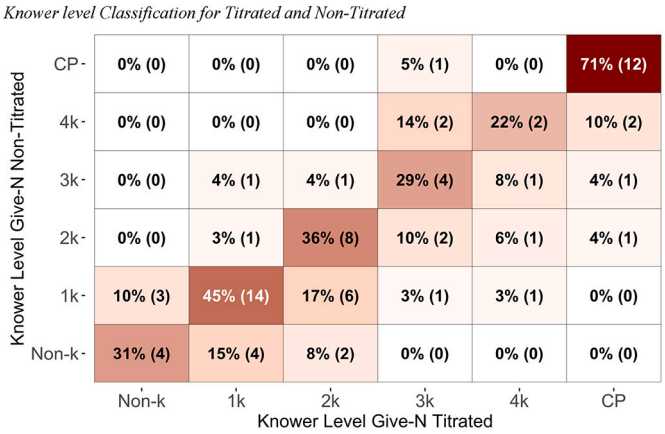

Fig. 3. Knower level Classification for Titrated and Non-Titrated Give-N.
Note. Titrated Give-N appears on the x axis, and non-titrated Give-N appears on the y axis. The percentages represent the percent effective agreement of both knower level assignments. Numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement.

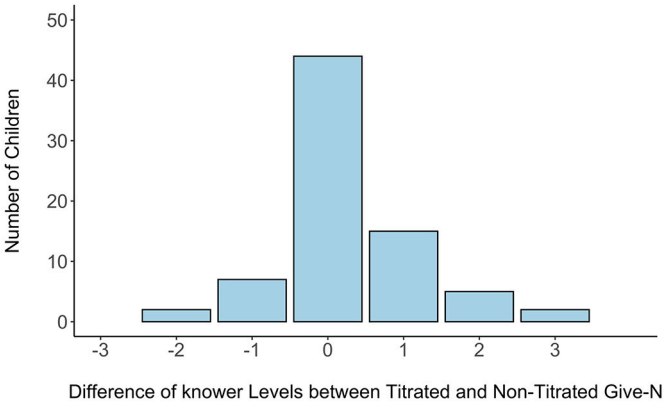

Fig. 4. Differences in Participant Knower Level Across Give-N Versions.
Note. The x-axis refers to the difference in participant knower level assignment between the titrated and non-titrated versions of Give-N. The 0 indicates no change in knower level assignment across the versions, while a positive number indicates a higher knower level assignment for titrated Give-N and a negative number indicates a higher knower level assignment for non-titrated Give-N. Amongst the 31 children with no matches between knower level assignments, 22 children had a higher knower level in the titrated version while 9 children had a higher knower level in the non-titrated version.

Table 9
Reliability measures and coefficients between the Titrated Give-N and WOC across different knower levels analyses.

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	44%	.55 *(w-l)* *95% CI,* .43 to .67 .71 *(w-q)* *95% CI,* .58 to .83 .30 *(unw)* *95% CI,* .18 to .43	.09 *(M)* *range:* .01 - .19	.11	.33 (*w-l*) *95% CI,* .16 to .50 .33 (*w-q*) *95% CI,* .10 to .56	.83 *95% CI,* .73 to .89
Subset-knower only	5 x 5	32% (effective)	.27 *(unw)* *95% CI,* .14 to .40	.11 *(M)* *range:* .01 - .19	*NA*	.30 *(unw)* *95% CI,* .16 to .44	*NA*
Non-knower vs others	2 x 2	20% (effective)	.28 *(unw)* *95% CI,* -.08 to .64	.84	.03	.79 *(unw)* *95% CI,* .60 to .91	*NA*
CP-knower vs others	2 x 2	46% (effective)	.52 *(unw)* *95% CI,* .29 to .74	.53	.04	.65 *(unw)* *95% CI,* .44 to .81	*NA*
Knower-level groups	3 x 3	72%	.45 *(w-l)* *95% CI,* .25 to .64 .52 *(w-q)* *95% CI,* .35 to .69 .40 *(unw)* *95% CI,* .20 to .61	.35 *(M)* *range:* .12 - .52	.07	.58 *(w-l)* *95% CI,* .42 to .74 .58 *(w-q)* *95% CI,* .38 to .78	.69 *95% CI,* .50 to .80

Note: The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

Table 7 shows the distribution of knower levels in the titrated and non-titrated Give-N tasks and for the WOC. On average participants could count to around 9 in the Highest Count task ($M = 9.1$) and their counting skills were variable (*range* = 0 to 30; $SD = 6.0$). Seventy children (93%) reached a knower level lower than their highest count across all 3 tasks.

4.2.1. Comparing titrated versus non-titrated Give-N

Table 8 provides a summary of the various coefficients assessing the relatedness of the titrated and non-titrated Give-N tasks. The analysis including all knower levels (0 to CP; Fig. 3) found that the relatedness of the two Give-N versions is high but lower than the relatedness of each respective version of the task to itself (in Experiments 1 and 2), suggesting that there are some real differences between the two versions above, beyond noise associated with test-retest reliability. This was also the case for the ICC statistic. As shown in Fig. 3, the degree of concordance between the tasks varies substantially across the individual knower levels. In particular, the concordance of 1-knowers and CP-knowers is especially high relative to other knower levels, resembling the reliability findings for Experiments 1 and 2.

To better understand how the two Give-N versions compared to each other, we next examined the differences in their outcomes (see Fig. 4). In total, there were 44 matches (59%) and 31 non-matches (41%). Overall, the non-titrated version generated significantly lower knower levels than the titrated version (Wilcoxon test rank test $V = 132.5$; $p = .02$). However, the majority of non-matches were differences of only one knower level ($n = 22$) as opposed to 2 knower levels ($n = 7$) or 3 knower levels ($n = 2$). Furthermore, when taking into account the Kappa statistics presented earlier, it appears that most children with non-matches were subset-knowers in both Give-N versions.

Next, we investigated whether there was an order effect. To do this, we performed a Wilcoxon rank test and found no significant effect of task order ($W = 2893$, $p = .76$). Regardless of Give-N type, from T1 to T2, a similar number of children increased their knower levels ($n = 17$) as decreased ($n = 14$).

4.2.2. Comparing the knower levels of Give-N titrated and WOC

Table 9 provides a summary of the various coefficients assessing the relatedness between the titrated Give-N and WOC. The analysis

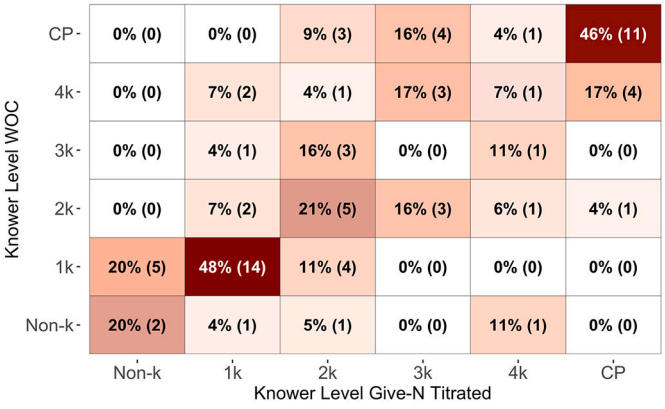

Fig. 5. Knower level Classification for Titrated Give-N and WOC.
Note. Titrated Give-N appears on the x axis, and WOC appears on the y axis. The percentages represent the percent effective agreement of both knower level assignments. Numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement.

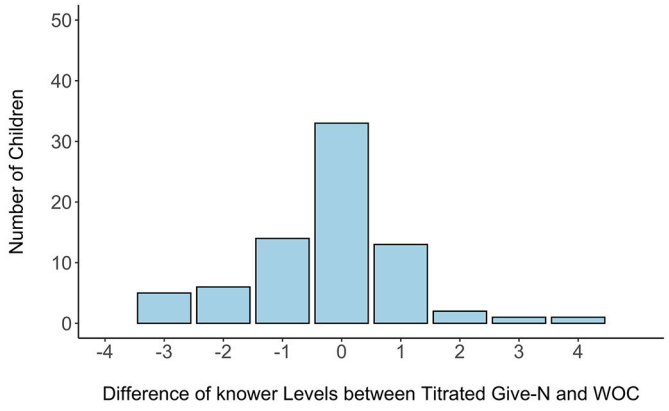

Fig. 6. Differences in Participant Knower Level Between Titrated Give-N and What's-On-This-Card.
Note. The x-axis refers to the difference in participant knower level assignment between the titrated version of Give-N and the What's-On-This-Card task such that 0 indicates no change in knower level assignment across the tasks, a positive number indicates a higher knower level assignment for titrated Give-N, and a negative number indicates a higher knower level assignment for What's-On-This-Card. There were 33 children with matching knower levels in the two tasks. Amongst the 42 children with no-matches between knower level assignments, 25 had a higher knower level in WOC and 17 in titrated Give-N.

including all knower levels (0 to CP; Fig. 5) found that the relatedness of the two tasks is acceptable but much lower than the relatedness between the two Give-N versions reported above. The ICC statistics corroborate those results. Fig. 5 also shows that concordance is high for 1-knowers and CP-knowers but very weak for 3- and 4-knowers. Fig. 6 shows the distribution of differences in knower levels between WOC and titrated Give-N. Overall, slightly more children were credited with a higher knower level in WOC ($n = 25$) compared to titrated Give-N (n = 17), though this difference was not significant (Wilcoxon test rank test $V = 325.5$; $p = .10$). Also, whenever there was a difference, most children presented a difference of 1 knower level ($n = 27$), as opposed to a difference of 2 ($n = 8$) or more levels (3 or 4 levels; n = 7).

4.2.3. *Comparing the knower levels of non-titrated Give-N and WOC*

Table 10 summarizes the various coefficients assessing the relatedness between the non-titrated Give-N and WOC. The analysis including all knower levels (0 to CP; Fig. 7) found that the relatedness of the two tasks is acceptable, similar to the outcome presented above of the assessment of WOC and titrated Give-N. This was also the case for the ICC statistic. The results are also similar in that the concordance was strongest for 1-knowers and CP-knowers but weaker for other knower levels. Fig. 8 shows that overall, more children were credited with a higher knower level in WOC ($n = 27$) compared to non-titrated Give-N ($n = 11$) and this difference was significant (Wilcoxon test rank test $V = 172$; $p = .003$). Interestingly, unlike in the case of the titrated task, the differences in knower levels between WOC and the non-titrated task tended to be larger; 21 children had knower levels that differed by one level while 17 participants had knower levels that differed by 2 levels ($n = 8$) or more (3 levels $n = 7$; 4 or 5 levels n = 2).

Table 10
Reliability measures and coefficients between the Non-Titrated Give-N and WOC across different knower levels analyses.

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	49%	.54 *(w-l)* *95% CI,* .41 to .67 .66 *(w-q)* *95% CI,* .51 to .81 .37 *(unw)* *95% CI,* .23 to .50	.08 *(M)* *range:* 0 - .19	.21	.39 (*w-l*) *95% CI,* .22 to .56 .39 (*w-q*) *95% CI,* .17 to .62	.80 *95% CI,* .65 to .88
Subset-knower only	5 x 5	40% (effective)	.34 *(unw)* *95% CI,* .20 to .49	.10 *(M)* *range:* .03 - .19	*NA*	.38 *(unw)* *95% CI,* .24 to .52	*NA*
Non-knower vs others	2 x 2	25% (effective)	.34 *(unw)* *95% CI,* .13 to .55	.80	.07	.76 *(unw)* *95% CI,* .57 to .89	*NA*
CP-knower vs others	2 x 2	39% (effective)	.45 *(unw)* *95% CI,* .23 to .67	.57	.08	.63 *(unw)* *95% CI,* .41 to .79	*NA*
Knower-level groups	3 x 3	71%	.41 *(w-l)* *95% CI,* .22 to .61 .46 *(w-q)* *95% CI,* .25 to .66 .38 *(unw)* *95% CI,* .18 to .59	.34 *(M)* *range:* .08 - .51	.13	.56 *(w-l)* *95% CI,* .39 to .73 .56 *(w-q)* *95% CI,* .36 to .76	.63 *95% CI,* .41 to .77

Note: The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

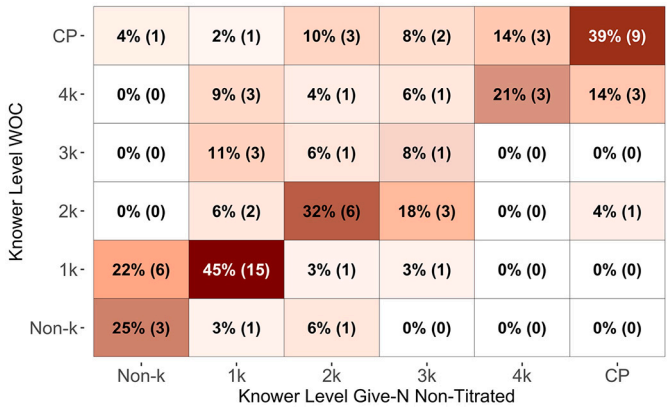

Fig. 7. Knower level Classification for Non-Titrated Give-N and WOC.
Note. Non-titrated Give-N appears on the x axis, and WOC appears on the y axis. The percentages represent the percent effective agreement of both knower level assignments. Numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement.

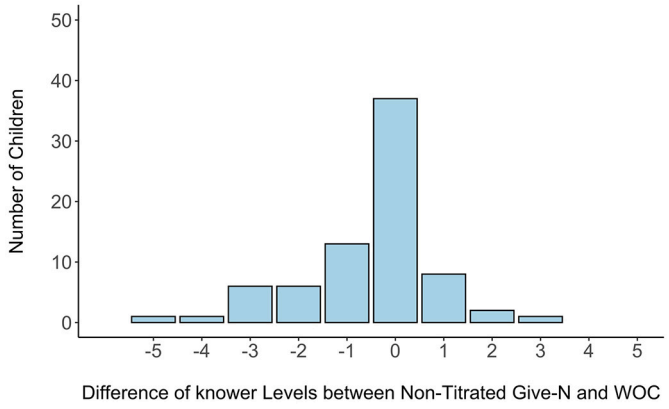

Fig. 8. Differences in Participant Knower Level Between Non-titrated Give-N and What's-On-This-Card.
Note. The x-axis refers to the difference in participant knower level assignment between the non-titrated version of Give-N and the What's-On-This-Card task such that 0 indicates no change in knower level assignment across the tasks, a positive number indicates a higher knower level assignment for non-titrated Give-N, and a negative number indicates a higher knower level assignment for What's-On-This-Card. There were 37 children who had matching knower levels across the two tasks and amongst the 38 who did not match, 27 had a higher knower level at the WOC task and 11, in the non-titrated Give-N task.

4.3. Discussion

In Experiment 3, we found that the relatedness of the titrated and non-titrated Give-N tasks was substantial, but that the titrated Give-N task generated slightly higher knower levels, typically 1 level greater than that of the non-titrated task. The comparison of WOC and the two Give-N versions found that the relatedness of WOC with either Give-N was acceptable, but not as strong as the relatedness of the two Give-N versions to one another. Interestingly, the Give-N versions differed in how they related to WOC. Although the titrated Give-N task did not generate systematically higher or lower knower levels than WOC (differences between outcomes were random), the non-titrated version produced significantly lower knower levels than WOC. This last result is consistent with previous studies in the literature suggesting that WOC attributes more knowledge of number words to children than Give-N (Baroody, Lai & Mix, 2017; Mou, Berteletti & Hyde, 2018; O'Rear, McNeil & Kirkland, 2020).

4.4. Post hoc analyses

As suggested by a reviewer, we conducted post hoc analyses on the relationship between children's counting abilities, age, and reliability. We conducted the analyses for subset-knowers and CP-knowers separately given the qualitatively higher degree of reliability amongst CP-knowers and non-knowers (making linear models difficult to interpret). One possibility is that highest count influences reliability (operationalized as concordance) for both groups, if it reflects children's executive functioning, attention, or ability to learn robust representations. An alternative possibility is that highest count may predict concordance only for CP-knowers, since only they can accurately count. With respect to age, predictions are more complicated, since amongst subset-knowers, lower knower levels exhibit higher reliability and children with lower knower levels tend to be younger. At the same time, older children should be less variable in how they respond relative to younger children. For CP-knowers, we expected that age might be positively related to concordance (because older children are better able to regulate their behaviors), or that it might be unrelated, given that CP-knowers have uniformly high levels of reliability.

To test these possibilities, we conducted two logistic regressions predicting Concordance in knower levels (yes/no) from Age and Highest Count. To maximize statistical power, we combined data from Experiments 1, 2, and 3 (only Give-N tasks). In our first model targeting only CP-knowers, Age ($\beta = -0.21$, $SD = 0.10$, $z = -2.09$, $p = .04$) and Highest Count ($\beta = 0.07$, $SD = 0.03$, $z = 2.16$, $p = .03$) were significant. Somewhat surprisingly, this effect of age was negative, suggesting that younger CP-knowers were slightly more likely to exhibit concordance than older CP-knowers. However, this effect was relatively small, and likely would not be found if substantially older CP-knowers were also tested. In our second model with subset-knowers (0-, 1-, 2-, 3- and 4-knowers), in which we added knower levels as a covariate, neither Age ($p = .98$) nor Highest Count ($p = .15$) were significant. Although the role of Age was not straightforward in this study, we believe that its role in influencing reliability should not be overlooked in developmental studies interested in the reliability of different tasks, as we discussed in our General Discussion. The results for Highest Count are consistent with our second prediction that counting abilities have an influence on concordance but only when children understand the purpose of counting and can use their counting skills in a task.

5. General discussion

In three studies we tested the reliability of the Give-a-Number task, while comparing two commonly used versions, the titrated and non-titrated versions. Overall, we found that the Give-N task is highly reliable, regardless of which version is used, though notable differences were found both between the tasks and across individual knower levels. First, in Experiment 1 we found that the titrated version of Give-N was very reliable overall, though the concordance of individual knower levels varied considerably, such that non-knowers, 1-knowers, 2-knowers, and CP-knowers exhibited fairly high concordance, while 3-, 4-, and 5-knowers did not.[11] We also found that the task could be reliably used to assign children to a less fine-grained tripartite classification of non-knower vs. subset-knower (1- through 4-knower) vs. CP-knower. Experiment 2 found almost identical results for the non-titrated Give-N task. In both experiments, testing location (either in-lab or off-site) didn't impact reliability. Finally, in Experiment 3 we tested the titrated and non-titrated versions within-subjects, and found that they exhibited a high degree of concordance, overall, although concordance was lowest for 3- and 4-knowers, similar to what was found when investigating test-retest reliability in Experiments 1 and 2. We also found

[11] Although their concordance was indeed low, 5-knowers were too infrequent (only 5 children ever obtain this classification) to draw firm conclusions from.

that while the overall distribution of knower levels was similar across versions, the titrated version produced significantly higher knower levels than the non-titrated task, though typically by just one knower level. Finally, although both tasks revealed differences from the What's-on-this-Card task, only the non-titrated task produced differences that were systematic (i.e., non-random) in nature. Just as it produced lower outcomes relative to the titrated task, the non-titrated task also produced lower outcomes than What's-on-this-Card.

Overall, these results support the continued use of Give-a-Number as a framework for classifying children, organizing findings, and predicting outcomes on other developmental measures. Also, our findings have both practical and theoretical implications regarding the use and interpretation of Give-a-Number in future studies. These implications relate to (1) the choice of task version in different research contexts, (2) how to use number knower levels to predict other outcomes in correlational designs, and (3) the validity of the knower level framework, as it relates to both the specific knowledge that is ascribed to children at particular levels by the different versions of the task, and also the status of higher, less reliable knower levels.

First, given our finding that both versions of the Give-N task generate relatively high degrees of reliability, the choice of which version to use should not hinge on reliability, but instead on secondary concerns of experimental design. On one hand, the titrated Give-N task features fewer trials, requiring less time, and does not require children who have relatively low knower levels to needlessly complete trials for large and unfamiliar numbers. For these reasons, it may be favored when Give-N is one of many tasks being administered, and when children are relatively young and unlikely to generate useful data for larger numbers. On the other hand, the titrated version of the task is considerably harder for inexperienced experimenters to learn and administer, since it requires adaptively changing the trial structure depending on children's individual behaviors, potentially increasing the likelihood of experimenter error. Also, because the titrated version does not systematically generate data for large numbers, it is not well suited to studies that seek to investigate how children respond to less familiar numbers (e.g., to test for knowledge beyond the child's knower level; Gunderson, Spaepen & Levine, 2015; O'Rear et al., 2020; Wagner & Johnson, 2011; Wagner, Chu & Barner, 2019), or that seek to conduct individual differences analyses, which generally assume that all participants have received the same measures (Geary, 2018; Geary et al., 2018; Geary, Vanmarle, Chu, Hoard & Nugent, 2019; Shusterman et al., 2016, 2017).

A second implication of this study concerns the use of Give-N to predict other developmental outcomes. Given the relatively high reliability of Give-N, our results suggest that it can be used in several different ways to fruitfully predict outcomes of other experimental measures, such as later mathematics achievement.[12] As noted in the Introduction, the use of a measure like Give-N to meaningfully predict other variables depends upon a relatively high test-retest reliability, since the strength of a correlation between any two variables is limited by the size of the correlation between the true value of the variables being measured, and the test-retest reliability of these measures taken individually. Therefore, in a study that attempts to correlate number knower level with another measure - e.g., a child's accuracy when making numerical estimates of dot arrays - the largest reliable correlation we might find between these measures would be limited by the reliability of Give-N (around 0.7) and the reliability of estimation accuracy (which is somewhat lower, around 0.57; Inglis & Gilmore, 2014). In this example, if the true correlation between these outcomes were 100%, then the highest detectable correlation would be 0.63 (i.e., 1 x $\sqrt{(0.7 \times 0.57)}$). This, in turn, has implications for the power required to detect reliable correlations, and thus for the size of the sample required for the study.

The third main implication of this study relates to the validity of the knower level system, and how individual knower level assignments should be interpreted. Across different studies using the Give-N task, researchers have often assumed, following Wynn (1992), that there are roughly five categories into which children might be classified - i.e., non-knowers, 1-knowers, 2-knowers, 3-knowers, and CP-knowers. However, some have allowed for the identification of higher levels, including 4-knowers and 5-knowers, and in some cases even higher. This approach is understandable, since it is possible that by restricting the possible subset categories to just three levels, researchers may underestimate the associative meanings that children acquire before they learn to accurately count and give large sets (and become CP-knowers). Our study, however, draws into question the interpretation of these higher knower levels. As we showed across three studies, whereas the non-knower, CP-knower, 1-knower, and 2-knower stages each individually exhibit high test-retest concordance, the 3-, 4-, and 5-knower stages are substantially less stable across sessions.

This apparent instability of higher knower levels is compatible with several interpretations. One possibility is that children at these higher knower levels are not actually 4- or 5-knowers, but instead are CP-knowers who attempt to count and make errors. Compatible with this, when we look at the three children from Experiment 3 who were classified as 5-knowers in the titrated Give-N, two of them were classified as CP-knowers in the non-titrated Give-N and one of them became a 4-knower. Similarly, one recent study by Krajcsi (2021) found that when children were prompted to fix Give-N errors by counting, this resulted in significantly more CP-knowers than when children were not prompted, or simply asked, "Is that N?" (see Le Corre et al., 2006, for related evidence). A second possibility is that children at higher subset levels aren't misclassified CP-knowers, but instead have noisy associative mappings between number words and approximate magnitudes. While some studies have argued for such a possibility (e.g., Wagner & Johnson, 2011), others have pointed out that such evidence is not robust once knower levels are assigned in keeping with Wynn's criteria, and when only those numbers clearly beyond the child's knower level are considered (e.g., Knower Level + 1; see Barner & Bachrach, 2010; Gunderson et al., 2015; Wagner et al., 2019; O'Rear et al., 2020). For these reasons, it is important in future work to not only assess whether children respond correctly on initial Give-N trials, but also (1) whether their initial response was the result of "grabbing" sets or an erroneous count (see Wynn, 1992), (2) whether they are able to fix their responses via counting when prompted, and (3) whether their overall pattern of responses for larger numbers is compatible with approximation, counting, or randomly guessing. Meanwhile, however, there is strong evidence that many children with higher knower levels are simply rare misclassifications of CP-knowers, and that when children are guessing noisily, this is restricted to the small number range (i.e., sets of 3–4 or less). Our work suggests that if children can be classified into higher subset-knower levels, these classifications are not reliable and should therefore be interpreted with caution. Future research should further explore this issue, and how the use of Give-N to identify higher subset stages might be validated.

The current study has several limitations that might be addressed in future studies. First, as in many studies of the Give-N task, the inferences permitted by our study is limited by sample size, which can impact estimates of reliability (Shoukri, Asyali & Donner, 2004; Sim & Wright, 2005). Ideally, in order to perform fine grained analysis of different subset-knower levels, one would want large numbers of participants categorized in each knower-level. However, because of their relatively low frequency, late subset-knowers can be particularly difficult to identify. For example, to obtain just 50 3-knowers, at a liberal rate of 8 per 100 children (based on our sample from Experiment 1), at least 500 children would need to be tested. Future studies might address this problem by combining the data collection efforts of multiple labs. A

[12] Note that in our study we have no evidence that the titrated or non-titrated version of Give-N would be more related to later learning, though other studies suggest that the non-titrated version may be more sensitive to small differences between children (e.g., O'Rear et al., 2020).

second potential limitation of this study is that sample characteristics (age ranges, cultural groups, socioeconomic groups, etc.) may impact reliability, leaving open the possibility that reliability may differ in different groups. For example, targeting children who progress through the knower level stages at a later age might result in higher reliability, if older children exhibit fewer random errors in performance. Similarly, the reliability of the CP- stage may be lower in cultures where children receive less training on counting routines than in the US (e.g., see Almoammer et al., 2013). Future studies should not assume that reliability will be identical across samples with different characteristics. A third limitation of our study is that we did not manipulate the time interval between the two Give-N tasks. For practical reasons, the Give-N tasks were administered in the same testing session, with only a brief counting task between administrations. Although we didn't find evidence for significant order effects, it is possible that reliability would be even greater with longer delays, given that the evidence for fatigue effects was slightly greater than evidence for improvement over the two sessions (in Experiment 1 and 2).

In summary, we found evidence that the Give-N task provides a useful framework for classifying children's number knowledge, and that it can be fruitfully used to explore correlations with other robust developmental phenomena. It will be important for future research to explore the impact of these findings on previously published work and to systematically examine the reliability of other tasks (e.g., WOC and Highest Count) frequently used in the literature. Given the widespread use of Give-N in the literature, future studies should also investigate the status of less reliable knower level stages, and their significance to theories of number word learning.

Data references

Authors: Elisabeth Marchand, Jarrett Lovelett, Kelly Kendro, David Barner.

Dataset 1: Data_Study1_Reliability.csv.
Dataset 2: Data_Study2_Reliability.csv.
Dataset 3: Data_Study3_Reliability.csv.
Repository: https://osf.io/48mke/
2021

Credit author statement

All authors (Elisabeth Marchand, Jarrett Lovelett, Kelly Kendro and David Barner) contributed significantly to this work.

Acknowledgements

This work received support from the Social Sciences and Humanities Research Council of Canada via a fellowship to E.M., a James S. McDonnell Foundation award to D.B., and a National Science Foundation award (ID: 2000827) to D.B. We would like to thank the participating children and families from Adventure Days Preschool, Seaside Preschool, Bright Beginnings Preschool, Chase Ranch Preschool, Gillispie School, Ridge City Preschool, Shelly Izzo's Daycare, St. Michael's Preschool, Northminster Preschool, Birch Aquarium, and Fleet Science Center. Special thanks as well to Samuel Beech, Anna Duran, Chris Fernandez, Hortesia Flores, Sonora Grimsted, Sara Lee, Emily Liu, Samuel Lucero, Ashlie Pankonin, and Kaithlyn Seifert. We also thank the members of the Language and Development Lab at UCSD, Attila Krajcsi and anonymous reviewers for their helpful feedback on this work.

References

Abreu-Mendoza, R. A., Soto-Alba, E. E., & Arias-Trejo, N. (2013). Area vs. density: influence of visual variables and cardinality knowledge in early number comparison. *Frontiers in Psychology, 4*, 805.

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., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology, 60*(1), 40–62.

Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology, 58*(2), 195–219.

Barner, D., Libenson, A., Cheung, P., & Takasaki, M. (2009). Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. *Journal of Experimental Child Psychology, 103*(4), 421–440.

Baroody, A. J., Lai, M. L., & Mix, K. S. (2017). Assessing early cardinal-number concepts. In *Proceedings for the thirty-ninth annual meeting of the North American chapter of the international group for the psychology of mathematics education* (p. 324).

Buelow, M. (2020). *Risky decision making in psychological disorders*. Academic Press.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423–429.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in Cognitive Sciences, 23*(10), 823–835.

Ceylan, M., & Aslan, D. (2018). Cardinal number acquisition of Turkish Children. *Journal of Education and e-Learning Research, 5*(4), 217–224.

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.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of cCinical Epidemiology, 43*(6), 551–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213.

Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General, 137*(1), 22.

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

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. Hoboken, NJ: Wiley-Interscience.

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

Geary, D. C., & Vanmarle, K. (2016). Young children's core symbolic and nonsymbolic quantitative knowledge in the prediction of later mathematics achievement. *Developmental Psychology, 52*(12), 2130.

Geary, D. C., Vanmarle, K., Chu, F. W., Hoard, M. K., & Nugent, L. (2019). Predicting age of becoming a cardinal principle knower. *Journal of Educational Psychology, 111*(2), 256.

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. (1993). A rational-constructivist account of early learning about numbers and objects. *Learning and Motivation, 30*, 61–96.

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.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorial in Quantitative Methods for Psychology, 8*(1), 23.

Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica, 145*, 147–155.

Jara-Ettinger, J., Piantadosi, S., Spelke, E. S., Levy, R., & Gibson, E. (2017). Mastery of the logic of natural numbers is not the result of mastery of counting: Evidence from late counters. *Developmental Science, 20*(6), Article e12459.

Krajcsi, A. (2021). Follow-up questions influence the measured number knowledge in the Give-a-number task. *Cognitive Development, 57*, Article 100968.

Krajcsi, A., Fintor, E., & Hodossy, L. (2018). *A refined description of preschoolers' initial symbolic number learning*. https://doi.org/10.31219/osf.io/2kh9s

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

Le Corre, M. (2014). Children acquire the later-greater principle after the cardinal principle. *British Journal of Developmental Psychology, 32*(2), 163–177.

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.

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.

Li, P., Le Corre, M., Shui, R., Jia, G., & Carey, S. (2003). Effects of plural syntax on number word learning: A cross-linguistic study. In *In 28th Boston University Conference on Language Development*. Boston: MA.

Luck, S. J., Cooper, H., Camic, P., Long, D., Panter, A., Rindskopf, D., & Sher, K. (2012). *APA handbook of research methods in psychology: Volume 1, foundations, planning, measures, and psychometrics*.

Marchand & Barner. (2019). The Acquisition of French Un. In *Proceedings of the 41st annual conference of the cognitive science society*.

Marušič, F., Žaucer, R., 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), Article e0159208.
Meyer, C., Barbiers, S., & Weerman, F. (2020). Many systems, one strategy: Acquiring ordinals in Dutch and English. *Glossa: A Journal of General Linguistics, 5*(1).
Meyer, D., Zeileis, A., & Hornik, K. (2021). *vcd: Visualizing Categorical Data. R package version 1.4-9*.
Moore, A. M., VanMarle, K., & Geary, D. C. (2016). Kindergartners' fluent processing of symbolic numerical magnitude is predicted by their cardinal knowledge and implicit understanding of arithmetic 2 years earlier. *Journal of Experimental Child Psychology, 150*, 31–47.
Mou, Y., Berteletti, I., & Hyde, D. C. (2018). What counts in preschool number knowledge? A Bayes factor analytic approach toward theoretical model development. *Journal of Experimental Child Psychology, 166*, 116–133.
Mussolin, C., Nys, J., Content, A., & Leybaert, J. (2014). Symbolic number abilities predict later approximate number system acuity in preschool children. *PLoS One, 9*(3), Article e91839.
Negen, J., & Sarnecka, B. W. (2012). Number-concept acquisition and general vocabulary development. *Child Development, 83*(6), 2019–2027.
Newman, A., Dickstein, R., & Gargan, M. (1978). Developmental effects in social facilitation and in being a model. *The Journal of Psychology, 99*(2), 143–150.
Nikoloska, A. (2009). Development of the cardinality principle in Macedonian preschool children. *Psihologija, 42*(4), 459–475.
Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.
O'Rear, C. D., McNeil, N. M., & Kirkland, P. K. (2020). Partial knowledge in the development of number word understanding. *Developmental Science, 23*(5), Article e12944.
Pfefferle, J. C., Machen, J. B., Fields, H. W., & Posnick, W. R. (1982). Child behavior in the dental setting relative to parental presence. *Pediatric Dentistry, 4*(4), 311–316.
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.
Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition, 123*(2), 199–217.
Purpura, D. J., & Simms, V. (2018). Approximate number system development in preschool: What factors predict change? *Cognitive Development, 45*, 31–39.
R Core Team. (2018). *R: a language and environment for statistical computing. R Foundation for Statistical Computing* (Vienna. http s. www. R-proje ct. org).
Rasmussen, E. E., Keene, J. R., Berke, C. K., Densley, R. L., & Loof, T. (2017). Explaining parental coviewing: The role of social facilitation and arousal. *Communication Monographs, 84*(3), 365–384.
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.
Sarnecka, B. W., & Lee, M. D. (2009). Levels of number knowledge in early childhood. *Journal of Experimental Child Psychology, 103*(3), 325–337.
Sarnecka, B. W., Negen, J., & Goldman, M. C. (2018). Early number knowledge in dual-language learners from low-SES households. In *Language and culture in mathematical cognition* (pp. 197–228). Academic Press.
Sarnecka, B. W., & Wright, C. E. (2013). The idea of an exact number: Children's understanding of cardinality and equinumerosity. *Cognitive Science, 37*(8), 1493–1506.
Schaeffer, B., Eggleston, V. H., & Scott, J. L. (1974). Number development in young children. *Cognitive Psychology, 6*(3), 357–379.
Schneider, R. M., Sullivan, J., Marušič, F., Biswas, P., Mišmaš, P., Plesničar, V., & Barner, D. (2020). Do children use language structure to discover the recursive rules of counting? *Cognitive Psychology, 117*, Article 101263.
Sella, F., Slusser, E., Odic, D., & Krajcsi, A. (2021). The emergence of children's natural number concepts: Current theoretical challenges. *Child Development Perspectives*. https://doi.org/10.1111/cdep.12428
Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13*(4), 251–271.
Shusterman, A., Cheung, P., Taggart, J., Bass, I., Berkowitz, T., Leonard, J. A., & Schwartz, A. (2017). Conceptual correlates of counting: Children's spontaneous matching and tracking of large sets reflects their knowledge of the cardinal principle. *Journal of Numerical Cognition, 3*(1), 1–30.
Shusterman, A., Slusser, E., Halberda, J., & Odic, D. (2016). Acquisition of the cardinal principle coincides with improvement in approximate number system acuity in preschoolers. *PLoS One, 11*(4), Article e0153072.
Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268.
Slusser, E., Ditta, A., & Sarnecka, B. (2013). Connecting numbers to discrete quantification: A step in the child's construction of integer concepts. *Cognition, 129*(1), 31–41.
Spaepen, E., Gunderson, E. A., Gibson, D., 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.
Stevenson, M., & Sergeant, E. (2021). *Package 'epiR'*.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics (Fourth S., editor)*. New York.
Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine, 37*(5), 360–363.
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, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition, 119*(1), 10–22.
Wagner, K., Chu, J., & Barner, D. (2019). Do children's number words begin noisy? *Developmental Science, 22*(1), Article 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.
Yantz, C. L., & McCaffrey, R. J. (2009). Effects of parental presence and child characteristics on children's neuropsychological test performance: third party observer effect confirmed. *The Clinical Neuropsychologist, 23*(1), 118–132.