Cross-task consistency and variability in assessing number word knowledge

Pierina Cheung[1], Rebecca Merkley[2], Theresa Wege[3], Sara Jasim[4], Daniel Ansari[5]

[1]National Institute of Education, Nanyang Technological University, Singapore

[2]Department of Cognitive Science, Carleton University, Canada

[3]Centre for Mathematical Cognition, Loughborough University, UK

[4]Department of Psychology, York University, Canada

[5]Department of Psychology & Faculty of Education, University of Western Ontario, Canada

Corresponding author: Pierina Cheung, cheung.pierina@gmail.com

Abstract

The cardinal meanings of the first few number words are often assessed with the Give-N Task, and children's knowledge of number words can be represented by "knower-level", with N-knower representing children who have acquired cardinal knowledge of number words up to N. In the current study, we sought converging evidence for knower-levels by examining the correspondence of knower-level classifications between the Give-N Task and the Point-to-X Task. In Study 1, we tested 69 preschool-aged children and found that children at the higher knower-levels often did not receive the same classification across tasks. Further, children who were classified as having acquired the cardinal principle on the Give-N Task were more likely than subset-knowers to count on the Point-to-X Task, but they did not count very frequently. In Study 2, we conducted secondary data analysis on an existing study that included a Point-to-X Task with different stimuli and more trial types. We again found that children's knower-level assessed with Give-N was not always associated with above chance performance for that number on the Point-To-X Task. These data suggest that knower-levels may adequately capture cardinal number knowledge for very small numbers but not for higher numbers. We discuss the practical implications of these findings and highlight a need for future studies to adopt a multi-method approach to establish a robust pattern of children's number word acquisition.

Keywords: number word knowledge, Give-N task, Point-To-X task, convergent validity, cross-task variability

How do we know what children understand about the meaning of number words such as “one”, “five” or “eight”? To study the cardinal meaning of number words, existing studies draw on methods akin to studies that examine the acquisition of other types of words, such as words for objects or properties of objects. For example, for young toddlers, researchers analyze where and for how long children look at different sets of objects on a screen upon hearing a number (Arias-Trejo et al., 2014; Kouider et al., 2006; Lukyanenko et al., 2016). For children aged 2 and up, children may be asked to point to or choose a set that matches a number word among some alternatives (Point-to-X Task and Match-to-Sample Tasks; Wynn, 1992; Le Corre et al., 2016; Shusterman et al., 2022; Silver et al., 2021; Slusser & Sarnecka, 2011), to say how many things there are in a set (How Many and What’s on this card? Tasks; Baroody et al., 2017; Gelman, 1993; Le Corre et al., 2006), or to produce a specific number of objects upon request (Give-A-Number Task; Wynn, 1990; Schaeffer et al., 1974).

Despite the various paradigms that have been reported in the literature, over the past couple decades, Wynn’s Give-N Task (1990, 1992, see also Schaeffer et al., 1974, Frye et al., 1989; also termed the Give-A-Number Task) has evolved to be the standard measurement for number word meanings (see Wege et al., 2022b, for a review). Typically framed as a game for giving toys to a puppet, children are asked to place different numbers of objects on a plate in this task. Using the Give-N task, Wynn has identified a stage-like progression of early number word learning. At first, young children can recite the count sequence, but cannot give the correct number of objects regardless of what number word is requested. Around the ages of 2 to 3, children begin to succeed on giving 1 object when asked for “one” while still failing on all other numbers. Over the course of several months, children then succeed on giving 2 when asked for “two”, and later on giving 3 when asked for “three”. This stage-like

progression continues till “three” or “four” (but see Krajcsi & Fintor, 2023). After learning the meaning of “three” or “four”, children appear to know that counting can be used to give all requested numbers on the Give-N task, typically up to “six” or “eight”. Based on this progression children can be classified into knower-levels. Children who fail to give any requested number and lack an understanding for how number words label numerosities of sets are labelled non-knowers. Subset-knowers know the meaning of a subset of the number words, and individually, they are termed 1-knowers, 2-knowers, 3-knowers, and 4-knowers (Wynn, 1990, 1992; Le Corre et al., 2006). Children who demonstrate an understanding for how counting connects number words to numerosities of sets are called Cardinal-Principle-knowers, or CP-knowers for short.

Why is the Give-N Task so popular? One reason is that previous studies have shown some evidence for the convergent and concurrent validity of knower-levels based on the Give-N Task. Convergent validity assesses whether knower-levels can be replicated across tasks, whereas concurrent validity assesses the relationship between knower-levels and other related aspects of numerical knowledge. In support of its convergent validity, first, Wynn’s original work has shown that children who showed knowledge of N on the Give-N Task also demonstrated knowledge of N on a two-alternative-forced-choice pointing task (the Point-to-X task, Wynn, 1992). Second, Le Corre et al. (2006) showed that children’s knower-levels from the Give-N Task generally matched with a task that had lower performance demands, i.e., What’s on this Card Task (What’s-On-This-Card), in which children are asked to say the number of items printed on a card. Third, they also found that only CP-knowers showed knowledge of cardinality when evaluating the counting behaviour of a puppet, and subset-knowers as a group did not.

In support of its concurrent validity, studies have shown a correlation between children's knower-levels on the Give-N Task and other numerical tasks. For example, Sarnecka and Carey (2008) found that CP-knowers know the successors of numbers such as "four" and "five", but subset-knowers do not. There is also evidence that the earlier children acquire the cardinal principle, the better they are on later math achievement (Geary et al., 2018; Chu et al., 2015). In addition, knowing the cardinal principle appears to be related to children's performance on non-verbal numerical tasks. For example, CP-knowers are better than subset-knowers at selecting the more numerous set (Abreu-Mendoza et al., 2013; Cheung & Le Corre, 2018; Negen & Sarnecka, 2015; Wagner & Johnson, 2011), and children who know more number word meanings are better at recognizing numerical equivalence between sets of objects (Mix, 1999a, 1999b, 2008a, 2008b). Collectively, these data provide a strong case for the knower-level framework. A recent study has further shown that the Give-N Task has acceptable test-retest reliability in classifying children into non-knowers, subset-knowers, and CP-knowers (Marchand & Barner, 2022). In short, evidence points to the Give-N Task being a reliable and valid measurement of number word knowledge.

However, recent studies have presented nuances of early number word acquisition. First, subset-knowers appear to have knowledge of number word meanings beyond their knower-level. Evidence for this comes from studies that extended the analytic approach typically used on the Give-N Task. For example, rather than classifying children as an N-knower, researchers have analyzed median responses on the next number (Wagner et al., 2019), average proportion correct on the next number (Barner & Bachrach, 2009), or crediting children with knowing N + 1 if they gave N + 1 correctly and *not* penalizing them for also giving N + 1 for other numbers (O'Rear, McNeil, & Kirkland, 2019). These analyses

show that some subset-knowers can sometimes correctly give the right number of objects when asked for the number beyond their knower-level. Second, some have argued that there are large number subset-knowers and raised questions about the knower-level framework. For example, by including higher numbers, a small group of children can be classified as 5-knowers, 6-knowers and so on (Krajcsi & Fintor, 2023; Posid & Cordes, 2018; Marchand & Barner, 2022). Based on this, Krajcsi and Fintor (2023) argued that the knower-level framework may not be an accurate description of children's number word acquisition.

Furthermore, the strongest evidence for both concurrent and convergent validity seems to apply to the CP-knower category, but not to the individual knower-levels within the subset-knower category. For concurrent validity, studies have demonstrated an effect of learning of the cardinal principle but findings are often not specific to knowing "one", "two", "three", or "four" (e.g., Cheung & Le Corre, 2018; Geary et al., 2018). For convergent validity, a closer look at Le Corre et al. (2006)'s study showed that 68% of children had the same knower-level classifications across tasks, and the discrepancies lie primarily within subset-knowers. Specifically, they found that while children who were classified as CP-knowers on the Give-N task were all CP-knowers on What's-On-This-Card, children who were 3- and 4-knowers on the Give-N task also showed knowledge of the cardinal principle on What's-On-This-Card (Le Corre et al., 2006; see also Cheung et al., 2021). In a more recent study, Marchand and Barner (2022) also did not find high correspondence between What's-On-This-Card and Give-N (both titrated and non-titrated versions of the task), with more than half of the children not receiving the same knower-level classification on both tasks. In particular, they found that among the individual knower-levels, the strongest correspondence was found for CP-knowers and 1-knowers, ranging around 45 to 70%

agreement for these two groups of children. A closer look at published data thus suggests that knower-levels from the Give-N Task do not always align with those from other tasks.

Differences in task demands may explain a lack of correspondence in knower-level classifications. For example, when there are mismatches in knower-levels, children tend to show poorer performance on the Give-N Task than on other tasks. On the Give-N task, children are asked to construct sets from a pile of objects, and their poorer performance could stem from coordinating the complex act of counting with remembering how many is requested. When task demands are lower such as the How Many Task in which children are shown a set of objects or pictures and asked how many, children tend to perform better (Baroody, Lai, & Mix, 2017; Mou, Zhang, Piazza, & Hyde, 2021; O'Rear et al., 2024). Nevertheless, in other tasks with lower task demands than Give-N, such as the What's-On-This-Card Task, children do not always perform better (Marchand & Barner, 2022).

Rather than using task demands to account for differences in children's performance between tasks, another possible explanation is that knower-levels are narrowly defined and may not adequately represent the underlying construct of cardinal number knowledge. Specifically, knower-levels are output classifications tied closely to the Give-N Task and are subject to scoring criteria of the Give-N Task. As a measurement of children's cardinal number knowledge, knower-levels are based on one type of behaviour – children's ability to generate a set that matches a number word – and may not provide a comprehensive understanding of the underlying construct of cardinal number knowledge. Knower-level classifications are also not standardized, with some studies applying the "N exclusively for N" rule and exclude children from being classified as knowing N if they give N for other numbers, but some do not (see Wege et al., 2022b). Different scoring criteria can affect

children's knower-level classifications. If knower-levels do not adequately capture cardinal number knowledge, then one may not expect high convergent validity of knower-levels between tasks that purport to measure children's cardinal number knowledge.

The present study

The goal of the present study is to seek convergent evidence for knower-levels as a measurement of cardinal number knowledge by examining the correspondence of knower-level classifications between the Give-N Task and the Point-to-X Task. We chose the Point-To-X Task because previous studies have yet to carefully examine Give-N and Point-To-X. For example, Wynn (1992) only included numbers surrounding a child's knower-level from Give-N, but did not examine a larger range of numbers. Silver et al. (2021) compared Give-N and Point-To-X Tasks and included numbers larger than N, but some pairs of comparisons do not assess whether children know N refers to exactly N because they could be solved by approximate number knowledge. For example, knowing that "two" refers to approximately two things would still allow children to answer correctly on 2 vs. 4 or 2 vs. 7 when asked to point to "two". Shusterman et al. (2022) included pairs of adjacent sets on the Point-to-X Task, but the target comparisons for N included pairs of sets in which N was paired with a smaller number, N – 1. For example, children were classified as knowing "two" if they chose a set of 2 for 1 vs. 2 and 2 vs. 3, but given the 2AFC nature of the task, if children only knew "one", they could still correctly choose 2 for 1 vs. 2 using mutual exclusivity. To fill these gaps, we designed a Point-To-X Task with pairs of adjacent sets that assess children's number word knowledge up to "six", which allows for a better comparison of knower-levels across tasks. Both tasks were matched on the scoring criteria for knower-levels. On the Point-

to-X Task, to be classified as knowing N, we only included pairs of comparisons in which N was paired with a higher number. If we find that children can be classified as an N-knower consistently across tasks, this suggests that the two tasks measure the same underlying construct, providing convergent evidence for individual knower-levels. However, if children cannot be classified consistently across tasks, this would provide weak evidence for the validity of individual knower-levels as measurements of the underlying cardinal number knowledge.

Study 1

Method

Participants

A total of 69 children (34 boys, 35 girls) who attended childcare centres in Singapore participated in this study. An additional three children were removed for showing a side bias on all trials (n = 1) or for not looking before choosing on all trials (n = 2). The average age of the sample was 45.8 months (32 months to 61 months, SD = 8.1 months). In a word document available on OSF <https://osf.io/gpbcm/?view_only=9e87158fda834735b62aebd439b9a68f>, we preregistered a target sample of 32 based on a kappa value of .7, which was computed using data from a previous study that compared two different measures for assessing knower-levels (Le Corre et al., 2006).[1] Nevertheless, we tested more children because there was an insufficient number of subset-knowers for in-depth analysis. We thus modified the sample size requirement with a revised target sample of having 10 children in each knower-level (non-

[1] We estimated our sample size based on a power analysis using Cohen's kappa. We computed the kappa value using Le Corre et al. (2006)'*s* Give-N vs. What's-On-This-Card data and found that kappa value was. 73 in their study. We estimated the distribution of the different knower-level groups based on a published study on bilingual preschoolers (Wagner et al. 2015), which was similar to our target sample in Singapore. We used their knower-level distribution for power analysis, using the kappaSize package in R. We collapsed across 3- and 4-knowers as one group. Formula: *CI5Cats(kappa0=0.7, kappaL=0.5, kappaU=NA, props=c(0.05, 0.1, 0.15, 0.2, 0.5), alpha=0.05)*

knowers, 1-knowers, 2-knowers, and 3-knowers) after reaching the target sample of 32.[2] We were unable to reach this revised target sample for the 1- and 3-knowers knowers when project funds ended. No data analyses were conducted before we officially ended data collection. The study was conducted in English, the primary language of instruction at preschools, and the sample included 75% of children who were exposed to languages other than English at home. Data were collected between 2019 and 2020 with additional data collected in 2022. All testing sessions were conducted in-person at the child's preschool.

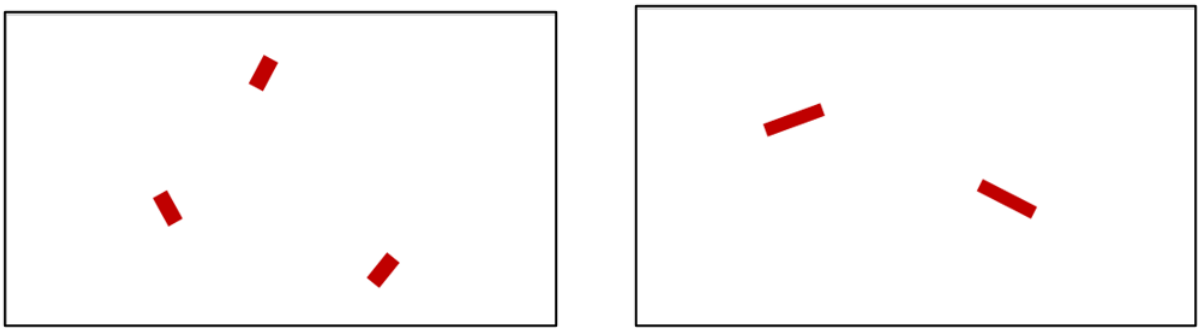
Tasks

Give-N. Children were asked to give a puppet N bananas on a plate, starting with "one". We followed the titration method. If the child succeeded, the experimenter asked for the next number, and if the child failed, the experiment asked for the number before it. The first trial was always "one", and the highest number asked was "six". After the child placed bananas on the plate, the experimenter asked, "Is that N?" to ensure that the child was done placing bananas. If the child responded no, the experimenter asked the child to place N bananas and repeated the "Is that N?" question. When the child responded yes, the experimenter lined up the bananas with spacing between each banana, and asked the child to count to check if there are N bananas. After the child counted, the experimenter asked "Is that N?" to assess whether children thought they had placed N. If the child responded no, they were asked to fix it to N. This procedure continued until the child responded yes to the "Is that N?" question after being asked to count and check.

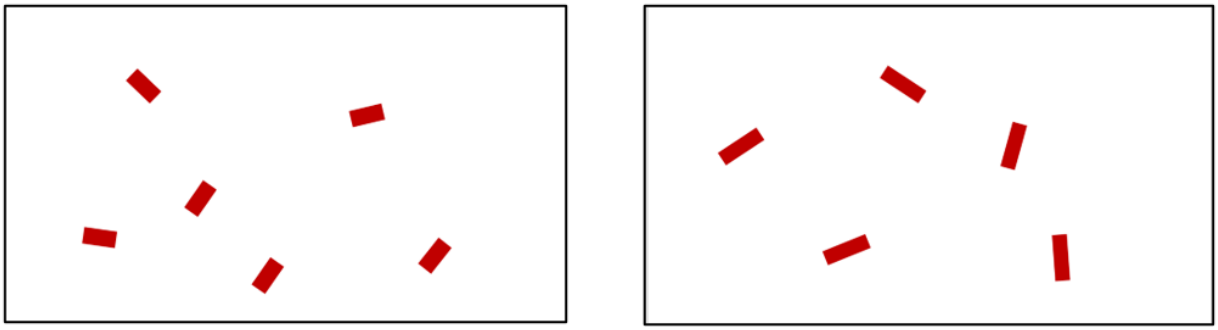
Point-to-X. Children were shown pairs of rectangles (e.g., 2 vs. 3) printed on paper and were asked to point to N blocks (e.g., "two" or "three"). There were six pairs of blocks: 1 vs. 2, 2 vs. 3, 3 vs. 4, 4 vs. 5, 5 vs. 6 and 6 vs. 7. For each pair, there were three trials asking

[2] We did not set this criterion for 4-knowers, because existing literature suggests that the sample of 4-knowers is typically much smaller than the other knower-levels.

for the smaller number and three trials for the larger number, for a total of 36 trials. For example, for 1 vs. 2, the experimenter asked children to point to “one” three times and “two” three times. The task was administered in two blocks: a small number block that included 1 vs. 2, 2 vs. 3, and 3 vs. 4, and a large number block that included 4 vs. 5, 5 vs. 6 and 6 vs. 7. The total surface area and perimeter were the same between the two sets in each pair. The stimuli set was drawn from a previous study on non-symbolic number comparison and was further adapted to create the comparison pairs for the current study (Cheung & Le Corre, 2018). Figure 1 presents sample stimuli from the study.



a) small number comparison: 2 vs. 3



2) large number comparison: 5 vs. 6

Figure 1. Example comparison pairs from the Point-to-X Task.

Procedure

The data collected for this study was part of a larger study on children's number word learning. Children were first tested on the Give-N Task and the Point-to-X Task, and those who were identified as subset-knowers on the Give-N Task were trained on learning the meaning of larger number words, which were administered in separate sessions after the Give-N and the Point-to-X Tasks. Additionally, children were always asked to count as high as they can after the Give-N Task. For a majority of the children (n = 52), the Give-N Task and the Point-to-X Task was counterbalanced. However, due to limited resources and a need for recruiting subset-knowers, the Give-N Task was always run before the Point-to-X Task for the remaining participants. Only data from the Give-N Task and the Point-to-X Task were reported here.

Coding

We determined a child's knower-level on the Give-N Task based on the following preregistered criteria. Children were classified as knowing N if they:

1. Correctly gave N objects 2 out of 3 times when asked for N
2. Failed to give the next number 2 times when asked for N + 1
3. Gave N objects no more than half as often for other numbers M

A child's highest N determined their N-knower status, and children who knew "five" or above would be classified as a CP-knower.

On the Point-to-X task, there were six comparisons (1 vs. 2, 2 vs. 3, 3 vs. 4, 4 vs. 5, 5 vs. 6), and each comparison was designed to assess knowledge of a particular number, from "one" to "six". For each comparison, three trials asked for the target number (e.g., for 1 vs. 2, "one") and three trials asked for N + 1 (e.g., "two"; see Table 1). Similar to Wynn's original design, a child was asked to choose N when N items were paired with N + 1 items; within the same comparison, children were also tested on whether they can apply knowledge of N to

rule out alternative sets, and we classified children as knowing N if they did *not* choose a picture of N items when asked for N + 1. Following the preregistered criteria, children were classified as knowing N on the Point-to-X task if they (see also Le Corre et al., 2016; Shusterman et al., 2022):

1. Correctly chose N items at least 2 out of 3 times when asked for N
2. Did not choose N items more than 2 times when asked for N + 1

A child who met these two criteria for N received a point for N, and the highest N determined each child's knower-level. Children who knew "five" or above were classified as CP-knowers.

Table 1. Study design for the Point-to-X Task.

Target Number (N)	**Target Number Comparison**	**Trials (Request Type: N and N + 1, N = target number)**
"one"	1 vs. 2	3 trials asking for "one", 3 trials asking for "two"
"two"	2 vs. 3	3 trials asking for "two", 3 trials asking for "three"
"three"	3 vs. 4	3 trials asking for "three", 3 trials asking for "four"
"four"	4 vs. 5	3 trials asking for "four", 3 trials asking for "five"
"five"	5 vs. 6	3 trials asking for "five", 3 trials asking for "six"
"six"	6 vs. 7	3 trials asking for "six", 3 trials asking for "seven"

Although not clearly stated in the preregistration document, we followed the Give-N coding and did not allow for gaps in children's number knowledge. We thus classified children as an N-knower where N was the highest number in a *consecutive* number sequence on the Point-To-X Task. For example, if a child received one point for knowing "one", "two", "three", and "five", this child would be coded as a 3-knower on the Point-to-X task. In

an exploratory coding, we allowed for one gap if the gap was followed by a consecutive number sequence. That is, if a child received one point for “one”, then “three” and “four”, we would code this child as a 4-knower. This exploratory coding method did not apply to children who knew up to “four” since there could not be a consecutive number sequence if the child did not receive a point for “five”. This exploratory coding changed the knower-level classification for four children, suggesting that most children who were coded as knowing N on the Point-To-X Task had knowledge of consecutive number sequences up to N.

There was no missing data in the small number block, but ten children did not have data on large number comparisons because an experimenter made an error (n = 6), children showed a response bias in the large number block (n = 3) or did not want to complete the large number block (n = 1).

Results and Discussion

Any deviations from the preregistration or exploratory analyses are noted in the manuscript. Multiple comparisons were corrected using Holm-Bonferroni. Knower-levels were double-coded manually in Excel, with all other analyses computed in R (R Core Team, 2018).

Preliminary analysis

We first explored whether there was a task order effect. Did children show better number knowledge when one task was administered before another? We tested this with the sample of children who had counterbalanced data and found no task order effects, regardless of whether the outcome was knower-levels on the Give-N Task ($F(1, 50) = 1.14$, $p = .29$) or knower-levels on the Point-to-X Task ($F(1,50) < 1$, $p = .9$). Knower-levels were entered as continuous variables in these analyses with task order as a dichotomous variable (Give-N first vs. Point-to-X first). Table 1 presents the mean knower-levels and SDs.

Table 2. Mean knower-levels and SDs by task order

	Mean knower-level on the Give-N Task (SD)	Mean knower-level on the Point-to-X Task (SD)
Give-N Task run first	3.36 (SD = 1.98)	2.91 (SD = 1.84)
Point-to-X Task run first	3.95 (SD = 1.75)	2.84 (SD = 1.86)

Preregistered analysis

Table 3 presents the number of children in each knower-level on the two tasks. Knower-levels were correlated between tasks (*tau* = .42, $p < .001$) but specific knower-level classifications corresponded only moderately, Cohen's *kappa* (weighted) = .41.

Table 3. The number of children in each knower-level on the Give-N Task and the Point-to-X Task.

	Give-N Task	Point-to-X Task
Non-knowers	11	7
1-knowers	6	14
2-knowers	10	12
3-knowers	9	15
4-knowers	3	4
CP-knowers[3]	30	17

[3] We found 1 5-knower on the Give-N Task and 5 children knew up to "five" but did not receive a point for "six" on the Point-to-X task. Following our pre-registered analysis plan, they were classified as CP-knowers.

Figure 2 presents a heat map comparing children's knower-levels across the two tasks. Overall, 32 children received the same classification across both tasks, including 2 (out of 11) non-knowers, 16 (out of 28) subset-knowers, and 14 (out of 30) CP-knowers. Most of the subset-knowers who received the same knower-level were 1-knowers or 2-knowers on the Give-N Task (13 of 16 1- and 2-knowers vs. 3 out of 12 3- to 4-knowers). These results suggest that knower-level classifications do not always align across both tasks, and the discrepancies may differ across knower-levels. Similar to Marchand & Barner (2022), we found higher concordance among 1-knowers and 2-knowers than those at higher knower-levels. Among those who did not receive the same classification (n = 37), 25 received a higher knower-level on Give-N than Point-To-X.

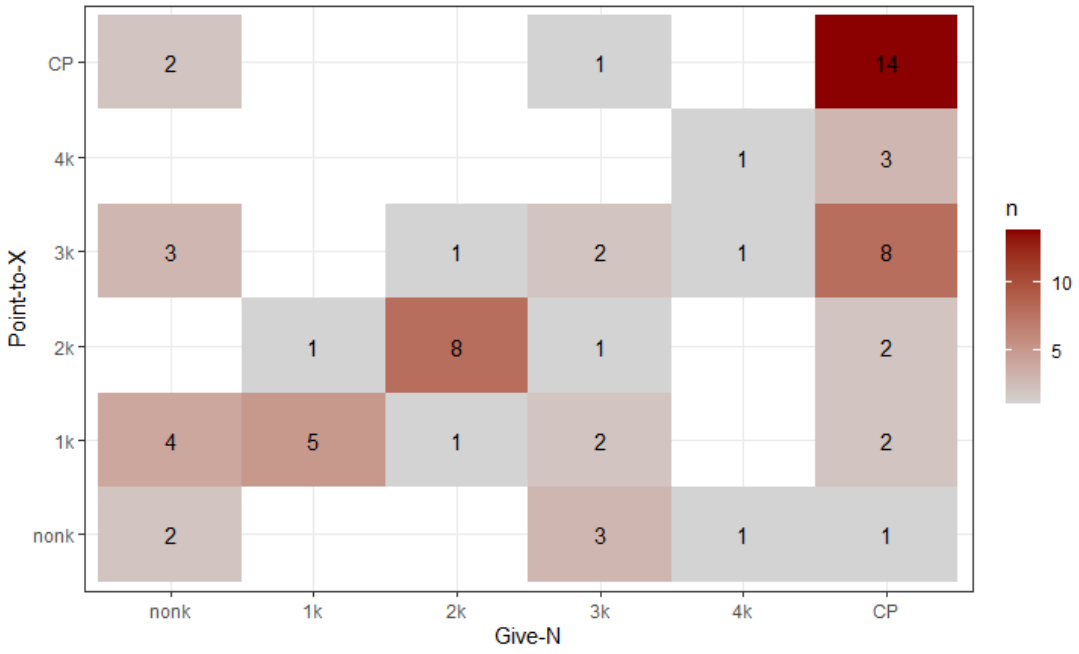


Figure 2. Heat map showing the correspondence of knower-level classifications between the Give-N Task and the Point-to-X Task. The numbers represent the number of children in that category.

Exploratory Analyses

The preregistered analysis showed that knower-levels across the two tasks were significantly correlated, but the correlations were not high. Also, children tend to have higher knower-levels on the Give-N Task than on the Point-To-X Task. One possibility is that some trial types on the Point-To-X Task are more difficult than others, and that may have masked children's knowledge. On the Point-to-X Task, for each target number comparison, there were six trials, with three trials asking for N and three trials asking for N + 1. For example, to assess children's knowledge of "one", children were presented with 1 vs. 2 and were asked to identify "one" and "two" across six different trials. That is, when shown 1 vs. 2, children were not only asked to identify a set of 1 when they heard "one", but they were also tested on whether they would restrict the label of "one" to the correct set and thus identify the other set when they heard a different number word, i.e., "two". Children were thus assessed not only on N, but also whether they can use knowledge of N to correctly identify N + 1 trials (see also Wynn, 1992). N-knowers could show poor performance because they lack knowledge of N + 1. We analyzed whether children's accuracy differed for N vs. N + 1. We focused on "known number comparison" on the Point-to-X Task, which was defined based on children's Give-N's knower-level -- i.e., 1 vs. 2 for 1-knowers, 1 vs. 2 and 2 vs. 3 for 2-knowers, 1 vs. 2, 2 vs. 3, and 3 vs. 4 for 3-knowers, and so on. CP-knowers were not included in this analysis. We performed linear mixed effects model with Request Type (N vs. N + 1) and Give-N Knower-Level (1-, 2-, 3- and 4-knowers) as predictors, with by-participant random intercepts (Figure 3). The analysis revealed no main effect of Knower-Level, ($\chi 2$ (3) = 7.23, p = .065), no main effect of Request Type ($\chi 2$ (1) = 0.36, p = .55) and no interaction, ($\chi 2$ (3) = 0.67, p = .88).

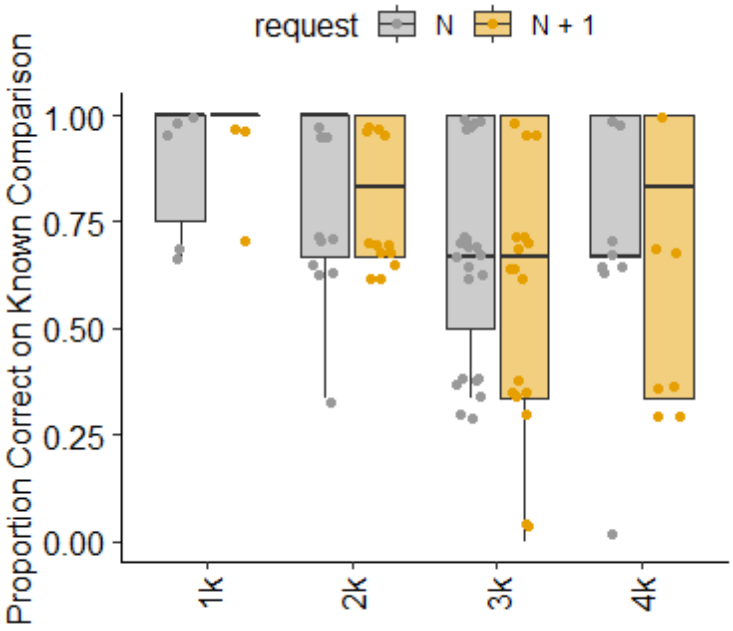


Figure 3. Boxplot and scatter plots showing children's performance on known comparison pairs on the Point-to-X Task.

Item-level analyses

To probe the correspondence of children's performance between the two tasks, we conducted item-level analyses on the Point-to-X task. Rather than coding children into different knower-levels, we computed proportion correct on the Point-to-X Task, and found a significant positive correlation between children's knower-levels on the Give-N Task and proportion correct on the Point-to-X Task, $r = .60$, $p < .001$.

We asked whether children classified as N-knowers on Give-N would show above chance performance on the known number comparisons and at chance on unknown number comparison on the Point-to-X Task. We excluded CP-knowers in this analysis. For example, for a Give-N 2-knower, 1 vs. 2 and 2 vs. 3 on the Point-to-X Task were considered this

child's known number comparisons, and 3 vs. 4 was the unknown number comparison. Due to the small sample, we excluded 4-knowers. We conducted one-sample t-tests with Holm-Bonferroni correction. We found that Given-N 1-, 2-, and 3-knowers were all above chance on their known number comparisons, t's > 3.08, p's < .015 (M_{1k} = .82, M_{2k} = .83, M_{3k} = .69). On the unknown number comparison, Give-N non-knowers were above chance, $t(10) = 4.69$, p = .003 (M_{non} = .83), but the other knower-levels were at chance, t's < 1, p's = n.s. (M_{1k} = .53, M_{2k} = .52, M_{3k} = .48). Figure 4 presents the item-level data by knower-levels and shows that the two tasks may be more aligned than suggested by the analyses based on knower-levels alone. We included 4-knowers and CP-knowers in the graph for comparison.

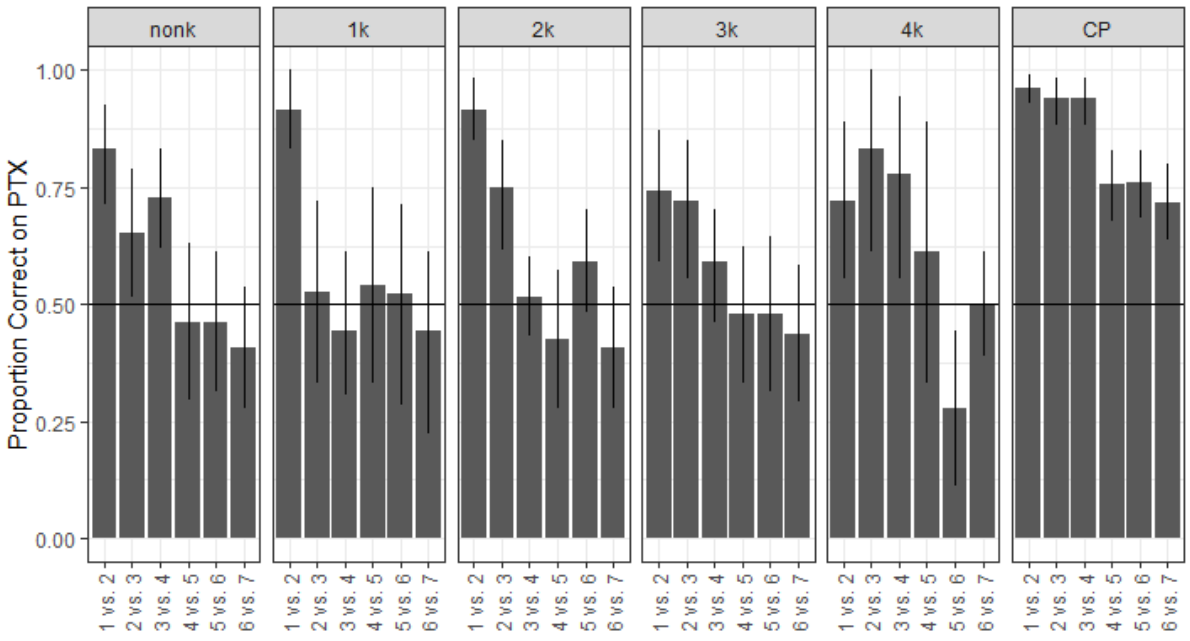


Figure 4. Children's performance on the Point-To-X Task at the item level as a function of knower-levels based on the Give-N Task. Error bars show a 95% CI interval around the mean.

Next, we conducted item-level analysis on children who were classified as CP-knowers on the Give-N Task and assessed whether they were able to select the correct set in each of the large number trials on the Point-to-X Task (i.e., 4 vs. 5, 5 vs. 6, and 6 vs. 7). We conducted one-sample t-tests with Holm-Bonferroni correction and found that they were significantly above chance on all large number comparisons ($t(29)$'s > 5.41, p's < .001, d's > 0.99). In addition, Give-N CP-knowers were significantly better than Give-N subset-knowers as a group on large number trials, $t(52)$'s = 4.77, p's < .001, d = 1.31 (M_{CP} = .76 vs. M_{subset} = .49; see Figure 2).

How often did children count on the Point-to-X Task?

Using another outcome measure, we examined whether Give-N CP-knowers were more likely to count to select the right number on the Point-to-X Task relative to Give-N subset-knowers. Using available data from videos (n = 42), we coded children's counting behaviour. Specifically, we coded whether children were overtly counting (e.g., counting aloud, with or without gestures), possibly counting (e.g., no overt counting can be heard from the video but children's lips were moving, their eyes were tracking the objects or they were pointing serially), no counting, and unclear cases. We focused on the large number block on the Point-to-X Task because previous studies show that Give-N CP-knowers and Give-N subset-knowers tend to differ from each other when asked for large sets on the Give-N Task (Le Corre et al., 2006). We combined overt counting and possibly counting, and found that on average, Give-N CP-knowers counted on 5.0 trials out of 18 trials on the Point-to-X Task (27.8%, SD = 7.0 trials) but Give-N subset-knowers never counted in the large number block on the Point-to-X Task (independent $t(40) = 3.34$, $p < .0018$, $d = 1.02$).

In sum, we compared children's performance on the Give-N and Point-to-X Tasks and found consistency and variability in knower-level classifications across tasks. In the knower-level analyses, we found better correspondence for 1- and 2-knowers, but poorer correspondence in the other categories. Item-level analyses reveal better correspondence for individual subset-knowers and CP-knowers. We also found that Give-N CP-knowers were more likely to count on the Point-to-X Task than Give-N subset-knowers, despite not counting very frequently on the Point-to-X Task. These findings show that children's behaviour and performance are comparable in some areas (e.g., analyzing performance as continuous rather than ordinal categories), but they also diverge in other areas (e.g., non-knowers' performance being better on the Point-to-X task, knower-level analyses). In Study 2, we further probed the lack of correspondence in the knower-level analyses using an existing dataset.

Study 2: Secondary Data Analysis on an Existing Dataset on Number Word Training

Study 1 showed that when the analyses focused on knower-levels alone, the Point-To-X Task did not always yield the same knower-level classifications as Give-N, and children generally scored lower on the Point-to-X Task. In Study 1, we examined request type (N vs. N + 1) as an alternative explanation, and did not find evidence that children's performance differed when asked for N vs. N + 1. However, the two tasks differed in other ways. For example, simple shapes were used in the Point-to-X Task but real objects were used in the Give-N Task. There were also more trials on the Point-to-X Task than on the Give-N Task, which could lead to fatigue and impact children's performance. The Point-to-X Task also had

only one trial type in Study 1, in which a target number was always paired with adjacent numbers. To further test cross-task variability, in Study 2, we conducted a secondary data analysis of a dataset that included the Give-N and Point-to-X Tasks (Wege et al., 2022a). In this study, there were fewer trials and pictures of familiar objects were used on the Point-to-X Task. There were also more trial types. We asked whether children identified as N-knowers on the Give-N Task demonstrated above-chance performance on corresponding number comparisons on the Point-To-X Task. This provided another test for task correspondence.

This study used secondary data collected by the same authors for another study (Wege et al., 2022a). Our primary question in the original study was to test whether children who saw number words referencing different kinds of objects would learn number word meanings (e.g., 2-knowers saw a set of three stars and a set of three hears) better than those who saw number words referencing the same kind of objects (e.g., different sets of three stars). We used a between-subject design to test this and we pre-registered children's performance on the Point-to-X task as the primary outcome measure. Additionally, we explored whether children were more likely to improve their knower-level on the Give-N task after the experimental training vs. control training (see Wege, 2022a, preprint). Here, we explicitly compared children's performance on the two tasks (Give-N and Point-to-X) in a secondary analysis, which was not conducted in the original paper. We used post-training data in this new analysis because the mechanism by which children learn number words is orthogonal to questions about the reliability or validity of the measurement of children's cardinal number knowledge.

Methods

Participants

We tested 65 children (M_{age} = 3;5, *Range* = 2;6 to 4;3; 30 girls), with 46 2-knowers and 19 3-knowers who were all English-speaking children. Children were recruited in London and Ottawa, Ontario, Canada in 2018 and 2019.

Original Study Design and Procedure

Children participated in a single-session training study that lasted for approximately 20 minutes, and half of them were assigned to the experimental training condition and the other half to the control condition. Prior to the number word training, children completed a titrated version of the Give-N Task to determine their knower-level and thus their trained number for the training session. The task was similar to that in Study 1. After determining the trained number, the experimenter assigned children to one of two training conditions. In both training conditions, children heard the trained number three times on each trial, for a total of 12 trials. The training conditions differed with respect to the type of number contrasts presented to children. In the experimental condition, children were presented with trained number sets that differed in the object kind (e.g., contrasting three yellow stars and three red hearts), whereas those in the control condition were presented with trained number sets that were of the same kind (e.g., three yellow stars vs. another set of three yellow stars; see Wege et al., 2022a, preprint). After training, they completed three tasks that assessed their number word knowledge, including the Point-to-X Task, the How Many Task, and the Give-N Task (see Figure 3 for study design).

Post training, we matched the total number of trials of all outcome measures such that both the Point-to-X Task and the Give-N Task included eight trials. On the Give-N Task post training, children were tested on their known number twice, their trained number three times,

and the number that was one higher than the trained number three times. The Give-N Task pre and post training thus differed from each other. The Point-to-X Task was administered on a touchscreen tablet. On the Point-to-X Task, children were shown two trials of trained number vs. known number (e.g., 2 vs. 3 for 2-knowers, and 3 vs. 4 for 3-knowers), three trials of trained number vs. the number after the trained number (e.g., 3 vs. 4 for 2-knowers and 4 vs. 5 for 3-knowers), and three trials of trained number vs. a set that was twice as large (3 vs. 6 for 2-knowers and 4 vs. 8 for 3-knowers). Regardless of trial types, the experimenter requested the trained number on all trials on the Point-to-X Task (i.e., for 2-knowers, the request was always "three"). The Give-N Task was administered twice (pre and post training), and the Point-to-X Task was only administered post-training.

Familiar objects were used on the Point-to-X Task, including pictures of stars, hearts, flowers, clouds, pebbles, bows, buttons, and crayons. On the Give-N Task, animal puppets and blocks (real objects) were used as stimuli.

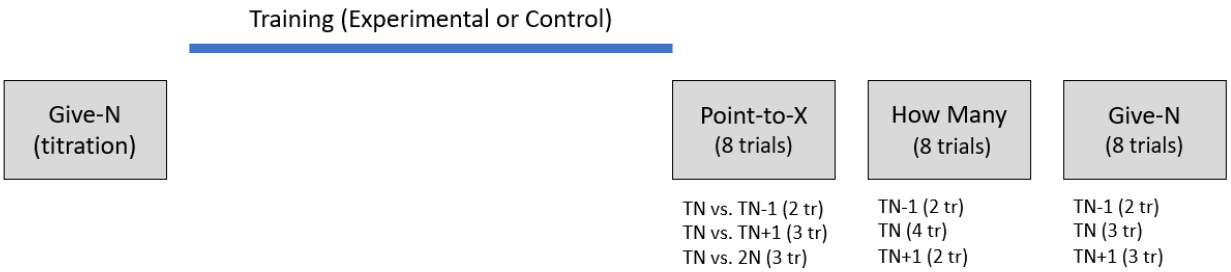


Figure 3 presents the experimental design for the study. TN refers to the trained number ("three" for 2-knowers and "four" for 3-knowers) and TN – 1 thus referred to N for N-knowers. The number of trials is presented in brackets.

Results and Discussion

Analysis plan

In our primary analysis published in Wege et al. (2022a, preprint), we asked whether there was an effect of training by comparing children's overall proportion correct on the Point-to-X task in the experimental and control training conditions (see also Huang et al., 2011). We also explored whether children's knower-levels were more likely to increase in the experimental than control conditions. We did not find a consistent training effect on both the Point-to-X Task or the Give-N Task. Importantly, we did not compare children's performance *across* the two tasks, which is the focus of the current analysis. In this study, we asked whether children who improved their knower-level on the Give-N Task after training showed corresponding improvement on the trained number comparisons on the Point-to-X Task. To address this, we first established that knower-level jumps were not due to measurement error of the Give-N Task. Then we conducted three analyses, first asking whether N-knowers were above-chance on known number comparisons on the Point-to-X Task, followed by two main analyses asking whether children who improved their knower-levels on the Give-N Task after training performed above chance on trained number comparisons on the Point-to-X Task.

Baseline analysis for interpreting knower-level improvements on the Give-N Task.

A total of 62 children completed the Give-N Task post-training. Overall, we found that 21 children (33.9%) improved their knower-level after training, including 15 2-knowers and 6 3-knowers, and 29 had the same knower-level. Twelve children had a knower-level drop, because they performed at 0% or 50% (out of 2 trials) on the known number trials. We first assessed whether knower-level differences observed post training reflect an effect of training or measurement error of the Give-N Task. As reported in the original study, we did

not find a consistent effect of training on knower-level improvements and it was likely due to our smaller than expected sample of subset-knowers (Wege, et al., 2022a, preprint). We also compared our data to those from Marchand and Barner (2022), who administered the titrated and non-titrated versions of the Give-N Task twice. We found that the proportion of children who showed knower-level jumps or drops, or demonstrated no change was significantly different between the current study and Marchand & Barner's reported data, *p's* < .004 (*chi-square* = 18.1 when comparing our study against Marchand & Barner's titrated Give-N data, and *chi-square* = 11.3 when comparing against their non-titrated data). The training study showed a larger proportion of children with knower-level jumps than Marchand and Barner (2022) (current study: 33.9% vs. Marchand & Barner's titrated Give-N study: 7.4% vs. Marchand & Barner's non-titrated Give-N study: 12.3%), but a similar proportion of knower-level drops (current study = 19.4%, Marchand & Barner =16.0% for both Give-N versions). These results suggest that children's knower-level jumps were unlikely due to measurement error.

New analysis comparing Give-N and Point-to-X

We examined whether children were above-chance on the known number comparisons on the Point-to-X Task, given that these comparisons included a known number (vs. the trained number, TN; TN vs. TN – 1; 2 vs. 3 for 2-knowers and 3 vs. 4 for 3-knowers). The request was always on the trained number ("three" for 2-knowers, and "four" for 3-knowers). These trials were thus the same as the N + 1 request trial on the Point-to-X Task in Study 1. We found that 2-knowers could reliably choose a set of 3 when shown 2 vs. 3, $t(45) = 3.93$, $p < .001$, $d = 0.58$ ($M = .72$, $SD = .38$), but 3-knowers could not reliably choose a set

of 4 when shown 3 vs. 4, $t(18) < 1$, $p = .42$, $d = 0.19$ ($M = .58$, $SD = .42$). These findings show that 2-knowers can use their known number knowledge (i.e., "two") to rule out alternatives but 3-knowers cannot (known number being "three").

For our main analysis, we asked whether those who improved on the Give-N Task demonstrated above chance performance on trained number comparisons (i.e., TN vs. TN + 1) on the Point-to-X Task. As a group, children who made knower-level jumps were at chance at selecting the trained number when it was paired with the next higher number, $t(22) < 1$, $p = .58$, $M = .54$.

These results reveal that children do not always show corresponding performance on the two tasks that are designed to tap into number word knowledge. It remains a possibility that the Point-to-X Task was difficult because pairs of comparison were of adjacent sets, and that children who made a knower-level jump may demonstrate corresponding improvement on the Point-to-X Task when the distance between sets was further apart. To test this, we asked whether children who made a knower-level jump demonstrated knowledge of the trained number on the Point-to-X Task when the comparison included a set twice as large. For example, children who learned "three" and were thus able to give 3 objects on the Give-N Task may find it easier to choose a set of three when shown 3 vs. 6 but not 3 vs. 4 on the Point-to-X Task. We analyzed the Trained Number vs. 2N trials on the Point-to-X Task, and found that children who improved on the Give-N task post-training remained at chance at this easier trial type, $t(22) < 1$, $p = .76$, $M = .52$. Thus, children who could give the trained number correctly on Give-N did not always correctly identify the trained number even when the sets were further apart on the Point-to-X Task.

General Discussion

The knower-level framework characterizes children's number word learning trajectory and has generated a rich body of literature on children's number development that extends beyond Wynn's seminal work (see Wege et al., 2022b, for a review and discussion). One of the basic premises of the framework is that being an N-knower means children have cardinal knowledge up to N. In this paper, we sought evidence for the convergent validity of knower-levels as a measurement of cardinal number knowledge by comparing children's knower-levels across two different measures, the Give-N Task and the Point-to-X Task. Across two studies, we found evidence that children's knower-levels assessed with Give-N were not always consistent with performance on the Point-To-X Task. First, in Study 1, we found that children who were classified as 1-knowers and 2-knowers tend to receive the same knower-level classification, but those at higher knower-levels, including 3-knowers, 4-knowers, and CP-knowers, did not. For example, children who could successfully give sets of up to 3 on the Give-N Task could not reliably choose sets of 3 from distractor sets on the Point-To-X Task. Second, children who could not produce a set of "one" on the Give-N Task often correctly identified small sets of objects on the Point-To-X Task. Third, we found that children who counted to give large sets of objects on the Give-N Task did not consistently count to identify the right number of objects on the Point-To-X Task. Finally, we assessed task correspondence using an existing dataset that included a Point-to-X Task that had fewer trials, used familiar objects, and had more trial types, and we found similar results. Children who have recently acquired number word meanings, which was measured by knower-level jumps on the Give-N Task, did not show corresponding improvement on the Point-To-X Task. These findings reveal that knower-levels may adequately capture cardinal number knowledge for very small numbers but not for higher numbers. Practically, our findings also suggest that the two tasks may not be used interchangeably if the goal is to classify children into different knower-levels.

Overall, we found that children received lower knower-levels on the Point-to-X Task than the Give-N Task. One possibility is that the Point-To-X Task is difficult and may not provide a sensitive test for the convergent validity of knower-levels. Given the nature of the Point-To-X Task, the target set is paired with a contrasting quantity, and in our study, each target number was paired with an immediate quantity (i.e., one lower or one higher). Children's ability to map number words onto a set of objects on the Point-To-X Task can be affected by the precision of their underlying number representations. Moreover, within each comparison, half of the trials included requests for a number larger than N (e.g., for 1 vs. 2, three trials asked for "one" and three trials asked for "two") and N-knowers may not be able to select the correct set when asked for N + 1. There were also more trials on the Point-to-X task, which could lead to fatigue in children. We have data to suggest that these explanations may not fully explain why Point-to-X Task generated lower knower-levels. In Study 1, we did not find that N-knowers performed worse when asked for N + 1 on the target number comparison. Study 2 included fewer trials and had contrasting sets that were twice as large as N, and we found that children who have acquired number word meanings, indicated by knower-level jumps on the Give-N Task, did not show corresponding improvement on the Point-to-X Task. Nevertheless, Study 2 was not designed to directly address alternative explanations and had a small sample size. It thus only provides weak evidence against the alternative explanations. Stronger designs are needed for understanding why children performed worse on the Point-to-X Task than on the Give-N Task. It remains possible that the Point-to-X Task underestimates children's knowledge of larger numbers because children do not recognize the need to count on a comparison task (Sophian, 1987, 1988, 1995).

Despite this, other studies have also reported knower-level discrepancies between tasks that assess number word knowledge (Marchand & Barner, 2021; Lee & Sarnecka, 2011; O'Rear & McNeil, 2019; Le Corre et al., 2006). A common approach to reconcile within-

child variability across tasks is to appeal to task demands. However, task demands may not fully explain knower-level discrepancies, because children do not always perform better on tasks with lower task demands (e.g., Marchand & Barner, 2022). It is also difficult to equate task demands. Thus, while task parameters may affect children's performance, the approach to explain away discrepancies may mask insights into what task demands may tell us about children's competence. As Sophian (1997) puts it, "task factors may sometimes obscure knowledge that children have; however, their impact is likely to be, at least in part, a function of what children know and how well they know it." (p. 286). Rather than appealing to task demands to explain cross-task variability within-child, we argue that data from this study suggest different use cases for the two tasks, and empirically, they raise questions about the validity of individual knower-levels as reflecting underlying cardinal number knowledge of specific numbers.

Using the Point-to-X Task in assessing number word knowledge

Our study revealed several interesting differences and similarities between the Point-to-X Task and the Give-N Task that have implications for number word assessment. First, we found that item-level analyses on the Point-to-X Task and knower-levels on the Give-N Task showed better concordance than analyses that focused on knower-levels alone. That is, we found that while children can give N on the Give-N Task, they do not consistently and reliably choose N when shown adjacent pairs of number, such that they may show above-chance performance but it was not consistent enough to be classified as knowing N according to knower-level criteria (i.e., at least 2 out of 3 correct on choosing N and did not choose N 2 or more times when asked for N + 1). Importantly, the level of agreement differs as a function of individual knower-levels. In Study 1, we found that Give-N 1-knowers and 2-knowers could reliably be classified as 1- and 2-knowers on the Point-to-X Task. These

children were able to identify a set of 1 from 1 vs. 2, and a set of 2 from 2 vs. 3, respectively, and they failed to identify larger numbers from higher number comparisons. The Point-to-X Task can thus identify children at the lower knower-levels, and interestingly, it reveals number knowledge for children who are otherwise classified as non-knowers on the Give-N Task.

Our data suggest that non-knowers are likely a heterogeneous group, and do not necessarily lack knowledge of number word meanings. As a group, they demonstrate knowledge of small numbers on the Point-to-X Task (see also Wagner et al., 2019). Previous studies have not explored what non-knowers understand because they are thought to lack number word meanings under the knower-level framework. Our findings suggest that their failure on giving numbers such as “one” or “two” on the Give-N Task could be related to coordinating a motor response in response to a number request or a lack of understanding of task instructions. Furthermore, diary studies showed that toddlers may have meanings for small numbers at an age that is preceding the use of the Give-N Task (Mix, 2009; Shusterman, Gibson, & Finder, 2010). These findings highlight a need for studying early number knowledge using a wider range of methods and exploring what non-knowers understand.

Our data also suggest that the Point-to-X Task is likely not a valid task for identifying children who would otherwise be classified as CP-knowers on the Give-N Task. Although we found that Give-N CP-knowers performed better at identifying large numbers and were more likely to count on the Point-to-X Task than Give-N subset-knowers, approximately half of these children who could generate sets labelled with “five” and “six” on the Give-N Task failed to reliably identify these numerosities on the Point-to-X Task. Also, children counted infrequently on the Point-to-X Task, making it difficult to reliably identify CP-knowers if a key difference between subset-knowers and CP-knowers is the understanding of the role of

counting. Previous studies have also found that 3- and 4-year-olds do not always recognize the role of counting in solving numerical problems such as comparing two sets of objects and creating numerical equivalence between two sets (Zhou, 2002; Sophian, 1987, 1988, 1995). These results suggest that there are limitations to CP-knowers' understanding of the role of counting, particularly in tasks that involve numerical comparisons.

Conceptualization of knower-levels

The lack of converging evidence for some of the individual knower-levels raises questions about what knower-levels represent and the underlying construct of cardinal number knowledge. Common across studies is the interpretation that knower-levels indicate cardinal knowledge up to N. That is, a 1-knower understands that "one" refers to sets of 1 but lacks knowledge that "two" refers to sets of 2, and a 2-knower understands the reference of "one" and "two", but not numbers higher than "two". Knower-levels are thus often treated as representing all-or-none knowledge of N. For example, in number word training studies, researchers trained N-knowers on the meaning of N + 1 (Huang et al., 2011; Gibson et al., 2019; Wege et al., 2022b). Children's performance on other number tasks has also been analyzed as a function of their knower-level, with numbers lower than or equal to N analyzed as "known numbers" and those beyond N as outside of their known number range (Schneider et al., 2021; Sokowlowski, et al., 2022). While these decisions could be interpreted as analytical and not theoretical in nature, our findings along with those from Marchand and Barner (2022)'s study suggest that we should caution against interpreting higher number knower-levels as representing all-or-none knowledge of N (Krajcsi & Fintor, 2023). Specifically, we found that 3- and 4-knowers' knowledge of number words might be different from those at the lower knower-levels. This is not the first study to demonstrate knower-level discrepancies with this group of children. In previous studies, 3- and 4-knowers do not

always show knowledge of N across repeated administration of the Give-N Task (Marchand & Barner, 2022) or across different number tasks (current study; Le Corre et al., 2006; Marchand & Barner, 2022; Orrantia et al., 2024). One possibility is that they may have begun to connect counting with the numerosities of sets (Cheung et al., 2021) and may thus show fluctuating number knowledge depending on their use of counting to give, fix an incorrect set, or identify sets (see Marchand & Barner, 2022, for related discussion). On this view, analyses based on children's accuracy on giving a set without considering other aspects of children's counting behaviour may not adequately capture cardinal number knowledge for higher numbers (Baroody, Lai, & Mix, 2017; Wege et al., 2022b). Previous studies have equated the operational definition of an N-knower with the underlying construct of cardinal knowledge of N (Sarnecka, 2015), and our findings suggest that focusing on children's ability to generate a set of objects may provide a narrow view on the development of cardinal number knowledge (see also O'Rear et al., 2024).

The knower-level framework has inspired a large body of work on early number acquisition, but open questions remain about knower-levels and the underlying representations of number word meanings (Sella et al., 2021). For example, what are the cognitive capacities that underlie small number word acquisition? What allows children to give small sets but not large sets correctly on the Give-N Task? Are small numbers really learned sequentially? The discovery of the pattern of number word learning may create an illusion that we have solved the puzzle of *how* children go through the number word learning trajectory, but many questions remain for the next decade of research on early number acquisition.

Limitations and Future Directions

One limitation of the current study is a lack of information about the reliability of the Point-To-X Task. It is possible that the Point-To-X Task, being a two-alternative forced-choice task, may subject to more measurement error than a production task such as the Give-N Task, because it has a wide range of task parameters, such as the choice of distractor sets, the kind of objects used, and the spatial arrangement of objects that could affect children's performance. Another limitation is that the current sample size is not suitable for estimating the likelihood of a child being classified as an N-knower on both the Give-N and Point-To-X Tasks. A third limitation is that the current study is not designed to assess the role of children's use of counting across the two tasks, because children were asked to count to check on the Give-N Task on every trial. Thus, children may perform better on the Give-N Task than on the Point-to-X Task. Still, we note that CP-knowers on the Give-N Task only counted 40% of the time on large number trials in the current study ("five", and "six"; see also Le Corre et al., 2006). An ongoing study is underway to address some of these limitations.

Moving forward, we suggest that future studies should adopt a multi-method approach to establish the pattern of children's number word acquisition. For example, future studies can include a wider range of number-word related tasks that assess the cardinal meaning of numbers (e.g., Give-N, What's-On-This-Card, Point-To-X, Match-to-Sample, Fast Cards, looking time measures) within-subjects and with a larger pool of participants to assess correlations between tasks. Longitudinal studies and microgenetic studies are also needed to fully understand the number word learning trajectory. The goal is to establish a more robust phenomenon of children's number word acquisition that can then be used as a basis for developing theories. Robust phenomena can constrain theory development and they should be "verifiable and detectable in several independent ways and not dependent on a specific theoretical framework or observation method" (Eronen & Bringmann, 2021, p. 780). The

approach to integrate among findings, rather than selecting among them, would provide an important step toward understanding children's early number word acquisition (see also LoBue et al., 2021).

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.

Arias-Trejo, N., Cantrell, L. M., Smith, L. B., & Canto, E. A. A. (2014). Early comprehension of the Spanish plural. Journal of Child Language, 41(6), 1356-1372Baroody, 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).

Cheung, P., & Le Corre, M. (2018). Parallel individuation supports numerical comparisons in preschoolers. *Journal of Numerical Cognition*, *4*(2), 380-409.

Cheung, P., Toomey, M., Jiang, Y. H., Stoop, T. B., & Shusterman, A. (2022). Acquisition of the counting principles during the subset-knower stages: Insights from children's errors. *Developmental science*, *25*(4), e13219.

Chu, F. W., & Geary, D. C. (2015). Early numerical foundations of young children's mathematical development. *Journal of Experimental Child Psychology*, *132*, 205-212.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779-788.

Frye, D., Braisby, N., Lowe, J., Maroudas, C., & Nicholls, J. (1989). Young children's understanding of counting and cardinality. *Child development*, 1158-1171.

Fuson, K. C. (1988). *Children's counting and concepts of number*. Springer.

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.

Huang, Y. T., Spelke, E., & Snedeker, J. (2010). When is four far more than three? Children's generalization of newly acquired number words. *Psychological Science*, *21*(4), 600-606.

Krajcsi, A., & Fintor, E. (2023). A refined description of initial symbolic number acquisition. *Cognitive Development, 65*, 101288.

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

Kachergis, G., Marchman, V. A., & Frank, M. C. (2022). Toward a "standard model" of early language learning. *Current Directions in Psychological Science*, *31*(1), 20-27.

Kouider, S., Halberda, J., Wood, J., & Carey, S. (2006). Acquisition of English number marking: The singular-plural distinction. *Language Learning and development*, *2*(1), 1-25.

Lukyanenko, C., & Fisher, C. (2016). Where are the cookies? Two-and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition*, *146*, 349-370.

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.

Lee, M. D., & Sarnecka, B. W. (2011). Number-knower levels in young children: Insights from Bayesian modeling. *Cognition*, *120*(3), 391-402.

LoBue, V., Reider, L. B., Kim, E., Burris, J. L., Oleas, D. S., Buss, K. A., ... & Field, A. P. (2020). The importance of using multiple outcome measures in infant research. Infancy, 25(4), 420-437.

Marchand, E., Lovelett, J. T., Kendro, K., & Barner, D. (2022). Assessing the knower-level framework: How reliable is the Give-a-Number task?. *Cognition*, *222*, 104998.

van Marle, K., Chu, F. W., Li, Y., & Geary, D. C. (2014). Acuity of the approximate number system and preschoolers' quantitative development. *Developmental science*, *17*(4), 492-505.

Mix, K. S. (1999a). Similarity and numerical equivalence: Appearances count. *Cognitive Development*, *14*(2), 269-297.

Mix, K. S. (1999b). Preschoolers' recognition of numerical equivalence: Sequential sets. *Journal of experimental child psychology*, *74*(4), 309-332.

Mix, K. S. (2008a). Children's equivalence judgments: Crossmapping effects. *Cognitive development*, *23*(1), 191-203.

Mix, K. S. (2008b). Surface similarity and label knowledge impact early numerical comparisons. *British Journal of Developmental Psychology*, *26*(1), 13-32.

Mix, K. S. (2009). How Spencer made number: First uses of the number words. Journal of Experimental Child Psychology, 102(4), 427-444.

Mou, Y., Zhang, B., Piazza, M., & Hyde, D. C. (2021). Comparing set-to-number and number-to-set measures of cardinal number knowledge in preschool children using latent variable modeling. *Early Childhood Research Quarterly*, *54*, 125-135.

Negen, J., & Sarnecka, B. W. (2015). Is there really a link between exact-number knowledge and approximate number system acuity in young children?. *British Journal of Developmental Psychology*, *33*(1), 92-105.

Odic, D., Le Corre, M., & Halberda, J. (2015). Children's mappings between number words and the approximate number system. *Cognition, 138,* 102-121.

O'Rear, C. D., & McNeil, N. M. (2019). Improved set-size labeling mediates the effect of a counting intervention on children's understanding of cardinality. *Developmental science, 22*(6), e12819.

O'Rear, C. D., McNeil, N. M., & Kirkland, P. K. (2020). Partial knowledge in the development of number word understanding. *Developmental science, 23*(5), e12944.

O'Rear, C.D. Kirkland, P. K., & Purpura, D. J. (2024). *The how many and give-N tasks: Conceptually distinct measures of the cardinality principle*. Early Childhood Research Quarterly, 66, 61-74.

Orrantia, J., Muñez, D., Sánchez, R., & Matilla, L. (2024). Mapping skills between symbols and quantities in preschoolers: The role of finger patterns. *Developmental Science,* e13529.

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.

Posid, T., & Cordes, S. (2018). How high can you count? Probing the limits of children's counting. *Developmental Psychology, 54*(5), 875.

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

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

Schröder, E., Gredebäck, G., Forssman, L., & Lindskog, M. (2022). Predicting children's emerging understanding of numbers. *Developmental Science, 25*(3), e13207.

Shusterman, A., Gibson, D., & Finder, B. (2010). Acquiring first number words: The developmental trajectory of children's meanings for "two.". In Proceedings of the 34th Annual Boston University Conference on Language Development [BUCLD 34] (pp. 375-384). Somerville, MA: Cascadilla Press.

Shusterman, A., Peretz-Lange, R., Berkowitz, T., & Carrigan, E. (2022). The development of early numeracy in deaf and hard of hearing children acquiring spoken language. *Child Development*, 93(5), e468-e483.

Silver, A. M., Elliott, L., Braham, E. J., Bachman, H. J., Votruba-Drzal, E., Tamis-LeMonda, C. S., ... & Libertus, M. E. (2021). Measuring emerging number knowledge in toddlers. *Frontiers in Psychology*, 3057.

Sokolowski, H. M., Merkley, R., Kingissepp, S. S. B., Vaikuntharajan, P., & Ansari, D. (2022). *Children's attention to numerical quantities relates to verbal number knowledge: An introduction to the Build-A-Train task*. Developmental Science, 25(3), e13211.

Sophian, C. (1987). Early developments in children's use of counting to solve quantitative problems. *Cognition and Instruction*, *4*(2), 61-90.

Sophian, C. (1997). Beyond competence: The significance of performance for conceptual development. *Cognitive Development*, *12*(3), 281-303.

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

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

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition*, *119*(1), 10-22.

Wege, T. E., Merkley, R., Cheung, P., Jasim, S., & Ansari, D. (2022a). What does ‘three’ look like? Analogical reasoning in early number word acquisition. PsyArXiv. https://doi.org/10.31234/osf.io/y57cp

Wege, T. E., Bourque, T., Merkley, R., & Cheung, P. (2022b). Thirty years of knower-levels: A systematic review of the Give-N task. PsyArXiv. https://doi.org/10.31234/osf.io/fh95s

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.