"Three yellow stars and three red hearts: Can subset-knowers learn number word meanings from multiple exemplars?

Theresa Wege
Centre for Mathematical Cognition, Loughborough University

Rebecca Merkley
Department of Psychology, Carleton University

Sara Jasim
Department of Psychology, York University

Daniel Ansari
Department of Psychology and Faculty of Education, Western University

Pierina Cheung
National Institute of Education, Nanyang Technological University

Abstract

Numerous studies have shown that number word learning is a protracted process. One challenge facing children learning the meaning of number word such as "one", "two", or "three" is that number words refer to a property of a set and not to individual objects. In this study, we focused on a sample of children who have not learned the meaning of small number words such as "two" and "three" and tested whether children could learn number words from examples of sets that help them focus on set size. Specifically, the experimental training condition included examples that highlight a common relational structure between sets through varying object properties in the sets (e.g., three yellow stars and three red hearts are both "three"), whereas the control condition did not vary object properties(e.g., two sets of three yellow stars with different spatial arrangement). We trained two- and three-knowers (N = 65) on the next number (i.e., three or four) and assessed their learning with a Two-Alternative-Forced-Choice task and Give-a-Number task. Overall, we found weak effects of training. We discuss our findings in the broader literature on number word learning and explore the possibility of analogical reasoning as a mechanism of number word learning.

Keywords: number word, subset-knowers, concept learning; training study; analogical reasoning

Introduction

One challenge in learning the meaning of number words is that number words refer to a property of a set (Bloom & Wynn, 1997; Merkley, Scerif & Ansari, 2017; Slusser & Sarnecka, 2011; Sullivan & Barner, 2011). Three blue pencils and three red chairs look nothing alike, yet these sets can both be labelled as "three".

To learn the meaning of number words, children thus have to attend to set size and ignore properties of individual objects such as shape or colour. Previous studies show that children may struggle with this. For example, when they see multiple sets of objects, they tend to pay more attention to object properties than the relational structure between sets (Christie & Gentner, 2010; Christie, Gentner, Call & Haun, 2016, Mix, 2008). For example, when asked to find the best match for a set of two yellow stars, most children would pick a set depicting stars or shapes that are yellow over a set depicting two objects. These results show that young children prefer to focus on individual object properties and not set size (Chan & Mazzocco, 2017; Mazzocco et al., 2020, Mix, 2008, Mix, 1999).

This preference for attending to individual object properties rather than to properties of setsmay explain why number word learning has been shown to be protracted. Compared to words for objects such as "ball" or words for properties of objects such as "red", children begin learning the meaning of number words (e.g., "one", "two", five") later. They begin to produce number words at around age 2, but they do not attach meaning to them. Initially, the count sequence is a meaningless string of words. Numerous studies have now shown that children from middle-income families in industrialized countries take 1 to 3 years to learn the meaning of number words. (e.g., Wynn, 1990, 1992; see Sarnecka, 2015 for review). These studies also show that the first few number words – "one", "two", "three", and "four" – are learned in sequence, one after another. And only after learning "four" do children appear to understand how counting represents number (e.g., Condry & Spelke, 2008; Le Corre et al., 2006; Wynn, 1990, 1992, among others). Children who know the meaning of the first few number words are called "N-knowers", where N refers to the number word that the child has acquired Collectively, N-knowers are termed subset-knowers, because they only know a subset of number word meanings. An N-knower can reliably generate sets denoted by the first few number words ("one" through "four") but fails to understand numbers higher than "four". For example, when asked to generate sets of objects, a two-knower can correctly generate a set of two objects but not a

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

set of three. A three-knower can correctly generate a set of three objects but not a set of four.

How do young children learn that "three" refers to a set of three things regardless of what those things are? Several studies have investigated how to aid this inference by showing children examples of sets labelled with a number word and applying a "count and label" procedure to highlight the meaning of number words as the last word of a count (e.g., one two three, *THREE* flowers). Despite extensive training that spans across multiple sessions, these studies tend to show that training children on the meaning of number words larger than their knower-level is difficult (e.g., Carey et al., 2017; Gibson et al., 2019; Huang et al., 2010; Mix, et al., 2012; O'Rear & McNeil, 2019). Exposing children to number contrasts as part of the training procedure (e.g., contrasting the trained number with a number less than that) also showed limited success (Gibson et al., 2019; Huang et al., 2010). For example, Gibson et al. (2019) found only approximately 35% of children improved their knower-level after four training sessions that included enriched number talk with number contrasts and counting training.

Theoretically, learning the meaning of small numbers is likely a distinct process from learning the meaning of counting (Barner, 2017; Carey & Barner, 2019). That is, training subset-knowers on the cardinal meaning of counting may involve a different process than training them on the meaning of individual number words. The former may involve training that highlight the counting procedures, whereas the latter may involve training that highlights set size as the reference of number words.

On the assumption that attending to set size – an abstract relational concept – is one challenge facing children learning number words, we hypothesize that number word learning may be facilitated if children are presented with examples of sets that shift their focus towards set size. For example, comparing a set of three red chairs with a set of three blue pencils highlights the common relational structure, namely that both are sets of three, irrespective of different properties of the individual objects in the sets. This contrasts with the training strategy in which a set of three is counted and labelled, with the last word emphasized. Previous studies tend to focus on counting as part of the training protocol. In this study, we tested whether highlighting set size by contrasting sets of different objects can help children learn number word meanings.

To do this, we adapted and improved upon a training paradigm previously used with subset-knowers (Huang et al., 2010). In the previous study, subset-knowers were presented with pairs of sets that contrast the trained number set with another set that differs only in set size (e.g., showing a two-knower cards depicting sets of identical looking dogs, one card with three dogs labeled as "three dogs" alongside another card with five dogs labeled as "not three dogs") (Huang et al., 2010). The training highlighted the reference of number word by showing negative evidence (i.e., cases in which the number word "three" did *not* apply vs. cases in which the label did apply). Using this paradigm, researchers found that children learned "three" in the context of trained examples but not to new examples (e.g., knowing how "three" applies to a set of dogs, but not how it applies to sheep). Even those children who successfully learned that number words can be used to refer to new examples only learned the *approximate* meaning of number words (e.g., "three" refers to a set of roughly 3 to 5 items) (see also Slusser, Stoop, Lo & Shusterman, 2017 and Posid & Cordes, 2018 for replications). These results suggest that subset-knowers can learn number word meanings from being shown contrasting sets, but children did not generalize it to sets outside of training and they did not learn the exact meaning of number words. In the current study, we improved upon this paradigm by providing examples that highlight a common relational structure between sets and support children in attending to set sizeas the reference of number words.

Current Study

Our goal was to test whether providing examples that highlight a common relational structure between sets can help young children learn the meaning of the first numbers. We focused on children who are just learning small number word meanings, i.e., subset-knowers, and trained them on the next number word (e.g., two-knowers on "three") in two between-subject conditions. In the experimental training condition, we showed children sets with varying object properties to highlight a common relational structure between sets (e.g., three yellow stars and three red hearts) and also set size contrasts (e.g., three yellow stars and four yellow stars) In the control training condition, comparisons and contrasts included only examples of sets with identical object properties (e.g., two sets of three yellow stars with different spatial arrangement, along with four yellow stars). We predicted that examples that highlight a common relational structure between sets may help subset-knowers infer that number words referred to set size independent of object properties. Therefore, subset-knowers who learned the next number from the experimental training should perform better than those who learned from a control training when asked to pick out the trained number from examples of sets.

Methods

The study was pre-registered on the Open Science Framework: https://osf.io/tafkp. Where deviations from the pre-registered protocol were necessary, we provide explanations below. All materials and data are available at: https://osf.io/w593e/

Participants

Children were recruited from preschools in London, Ontario and Ottawa, Ontario, Canada. A total of 149 children participated, but only children who were two- or three-knowers were included in the study, following Huang et al. (2010). Children were pseudo-randomly assigned to the experimental or control training conditions, to ensure similar numbers of children of each knower-level in each training condition. Our final sample included 65 children, with 31 in

the experimental condition (age in years;months: *M* = 3;5, *SD* = 0;5, *min* = 2;7, *max* = 4;0, 2-knowers: *n* = 21) and 34 in the control condition (age in years;months: *M* = 3;5, *SD* = 0;6, min = 2;6 , max = 4;3, 2-knowers: *n* = 25). There were 35 boys and 30 girls. All children were fluent in English. We pre-registered a target sample of 72 children (36 in each training condition) based on a power analysis, but data collection was terminated early because recruitment resources were exhausted.

Tasks

Children first completed the Give-a-Number task (Wynn, 1992) to determine their knower level, and then participated in a brief exposure training on the next number word, and check trials to assess whether they hadlearned number word meanings in the context of trained examples. After this, children participated in three post-training tasks in the following order: a Two-Alternative-Forced choice task, a How Many task and another Give-a-Number task. The Two-Alternative-Forced choice task was preregistered to be the primary outcome measure, similar to Huang et al. (2010). As an exploratory measure, we compared children's knower-level on the pre vs. post-training Give-a-Number task. The data from the How Many task are not reported here.

Coding knower-levels: Give-a-Number Children were asked to give a puppet sets of different numbers of objects. Children were shown a monkey, giraffe, and tiger puppets and asked to choose an animal to play with. They were told to give Mr. [animal] different numbers of blocks because he wants to play with them. Ten blocks, which were identical in size and colour, were arranged in a pile in front of the child. For each trial, the experimenter asked, "Can you give Mr. [animal] *x* blocks?" After the child finished putting blocks in front of the puppet, the experimenter asked, "Is that *x*?" If the child said no, the experimenter asked, "Can you make it *x*?" Once the child said yes, the experimenter recorded their response and moved on to the next trial. We used the titration method so the number of objects asked for was adjusted based on children's responses. The experimenter asked for the next number if the child gave correctly, or one number lower than *x* if the child gave incorrectly. Children were first asked to give two objects to the puppet. The task ended when children responded incorrectly twice for any number or when they responded correctly twice for the number eight. A child was considered a knower of a number if they correctly gave that number twice, and failed to give the next number twice. Scoring did not consider whether knowers also gave known numbers on trials asking for other numbers. Only two- and three-knowers were included and were analyzed as a group as stated in our pre-registration.

Brief Exposure Training Children were shown examples of sets labeled with the trained number word. Two-knowers were trained on the number word "three" and three-knowers were trained on "four". In the following sections, we refer to the trained number word as *n*, meaning that *n-1* is the child's knower level. In each training trial, children saw three example sets that were presented sequentially: a set of *n* objects, another set of *n* objects as a comparison and a set of *not-n* objects as a contrast. Four types of objects were used in the example sets: yellow stars, red hearts, green flowers, and blue clouds. We designed two training conditions: an experimental training designed to support inference about set size by highlighting a common relational structure and a control training. In both the experimental and control conditions, the initial set of *n* objects and the set of *not-n* objects that was used as a contrast contained the same objects. The only difference between the two conditions was that the set of *n* objects that was used as a comparison contained different objects (Figure 1B) in the experimental condition but the same objects in the control condition (Figure 1A). Thus, in the control training, the same objects were used in all three example sets.

The training was presented on a tablet held by the experimenter and children were offered a puppet to hold during the training. They were told that the puppet wanted their help to learn numbers. There were three blocks of training, with four training trials in each block that presented examples of sets of *n* for each of the four objects. There was a total of 12 training trials. The same trial structure was used for both the experimental and control training (Figure 1). In each trial the three example sets were presented in the following way:

1. Example: A set of the trained number *n* depicting one of the four possible objects (e.g. yellow stars) appeared in the middle of the screen and remained on the screen throughout the trial. The experimenter labeled the example set: "Look, [*object*]! There are [*n*]!"
2. Comparison: Another set of the trained number *n* appeared either to the left or the right of the first example set remained on the screen. The side was counterbalanced across trials. The experimenter said: "Look, [*object*]! This is also [*n*]!"
3. Contrast: A final set which did not depict the trained number n appeared on the other side of the screen. Across trials, it was counterbalanced whether this contrasting set depicted *n-1*, *n+1*, or *2n*. The experimenter said: "Look, [*object*]! There are not [*n*]!"

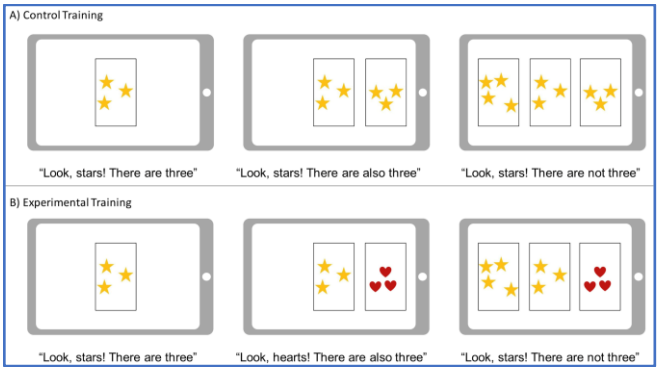

Figure 1: Example sets and training procedure in a single trial for A) Control Training and B) Experimental Training.

Training Check Immediately following the training, children completed four trials to check whether they were able to select sets containing *n* among the trained examples. On each trial, children were shown two sets of the objects used in the training and asked to select the set that contained the trained number ("Look, [*object*]! Which picture has [*n*]? Point to the picture that has [*n*]!"). The distractor sets contained *n-1*, *n+1*, *2n*, *n-1* objects. We preregistered to include children who correctly responded to at least three out of four training check trials.

Two-Alternative-Forced-Choice (2AFC) Similar to the training check trials, children were shown two sets and asked to select the set that contained the trained number *n*: "Look, [*object*]! Which picture has [*n*]? Point to the picture that has [*n*]!". The distractor sets depicted *n-1*, *n+1* or *2n*. There were two trials for *n* vs. *n-1* and three trials each for *n* vs. *n+1* and *n* vs. *2n*, for a total of eight trials. The sets did not contain the same objects as in the training. Half of the trials presented sets with familiar objects that belong to the same kind as the objects in the training but were not identical (e.g., yellow stars that had a different texture and shading than the yellow stars used in training). The other half of the trials presented sets of novel objects children had not seen during training (grey pebbles, pink bows, turquoise buttons, purple crayons). The order of trials was randomized.

Give-a-Number The task instructions were the same as described above, but a fixed trial list was used instead of the titration method. Children were again asked to give objects to a puppet. They were asked to give *n* items three times, *n +1* items twice and, *n-1* items twice for a total of eight trials. We used this task in our exploratory analysis and coded children who gave the trained number *n* correctly at least two out of three times as having advanced their knowledge of the meaning of the trained number word. Out of the full sample of 65 children, 13 children did not complete this post-training Give-a-Number task due to fatigue or interruption of the testing session.

Results

Preliminary Analyses: Training check & Inclusion

The pre-registered inclusion criterion was for a child to answer at least three out of four trials correctly on the training check trials. Average performance on these trials was lower than expected, and only a small subsample of children reached the inclusion criterion ($N = 25$ out of 65). To increase statistical power, we also conducted the main analysis on the full sample of children and conducted all exploratory analyses on the full sample. On average, the full sample performed significantly above chance on the training check trials, $M = 61.5\%$, $SD = 25.4\%$, $t(64) = 3.66$, $p < .001$. We reported all preregistered main analyses for the sample based on the pre-registered inclusion and additionally for the full sample. We combined two- and three-knowers in all our analyses.

Confirmatory Analyses: Experimental training vs. control training on 2AFC accuracy

Our main question was whether children who received the experimental training performed better than those who received the control training when asked to pick out the trained number in a 2AFC Task. Figure 2 presents children's average accuracy on the 2AFC Task by training conditions. For the subsample that met the pre-registered inclusion, children who received the experimental training ($N = 15$, $M = 70.0\%$, $SD = 19.4\%$) had higher accuracy than those who received the control training ($N = 10$, $M = 61.3\%$, $SD = 23.2\%$), but this difference was not statistically significant, $t(23) = 1.024$, $p = .158$, $d = .482$, $BF_{10} = .876$. For the full sample, children who received the experimental training ($N = 31$, $M = 59.3\%$, $SD = 21.2\%$) had a higher accuracy than those who received the control training ($N = 34$, $M = 49.3\%$, $SD = 20.4\%$), and this difference was statistically significant, $t(63) = 1.941$, $p = .028$, $d = .482$, $BF_{10} = 2.363$.

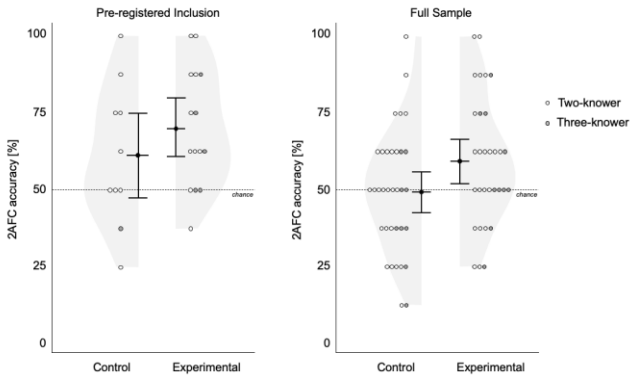

Figure 2: Children's performance on the Two-Alternative-Forced-Choice task compared between training conditions. Left: pre-registered inclusion after removing children who did not pass the check trials, right: full sample without removing children who did not pass the check trials. Black points and error bars represent Mean and 95% Confidence Interval

Exploratory Analyses

Exploratory analyses were conducted on the full sample only. We first tested whether children learned that the trained number refers to exactly *n* on the 2AFC Task. If the experimental training facilitated children's learning that *n* refers to exactly *n* and not approximately *n*, we would expect significantly better performance on the n vs. *n+1* trials relative to the control condition. We conducted an independent samples t-test comparing the two conditions specifically on the n vs. n+1 trials. We did not find support for this prediction. In the full sample, children who received the experimental training ($N = 31$, $M = 55.9\%$, $SD = 32.6$) did not perform significantly better than those who received the

control training (N = 34, M = 43.1%, SD = 30.2%), $t(63)$ = 1.640, p = .053, d = .407, BF_{10} = 1.467.

In a second exploratory analysis, we tested whether children's success at picking out the trained number was mainly driven by accuracy on sets familiar from the training. We found no evidence for a difference in accuracy between trials with novel or familiar sets in either training condition (main effect of set type: F (1,63) = 0.119, p = .731, BF_{10} = 0.193; interaction between set type and training condition, F (1, 63) = 0.549, p = .461, BF_{10} = 0.347).

Finally, we tested whether children in the two training conditions had learned the meaning of the next number word by analyzing whether they advanced their knower-level on the Give-a-Number task. Out of the full sample, 52 children completed the post-training Give-a-Number task. Over half of the children who received the experimental training (N = 15/27, 56%) gave the trained number correctly at least in two out of three trials, compared to only about a third of the children who received the control training (N = 8/25, 32%). This difference was not statistically significant, x^2 = 2.920, p = .087, BF_{10} = 1.373.

Discussion

In this study, we investigated how children learn the meaning of number words from examples of sets. Previous studies looking into early number learning from examples have aimed to investigate the role of domain-specific mechanisms such as counting (e.g., Carey et al., 2017; Mix, 2012; O'Rear & McNeil, 2019; Posid & Cordes, 2018). In the current study, we tested whether examples that highlight a common relational structure between sets can facilitate subset knowers' learning of the next number. Critically, unlike previous studies, we designed an experimental training paradigm that included different examples of sets that can be labelled by the same number word (e.g., three yellow stars and three red hearts), in addition to sets of different set size but otherwise had the same objects (e.g., three yellow stars and four yellow stars). We did not find evidence for a training effect in the pre-registered analysis that included children who passed the training check. We did find a difference in training conditions in the full sample, when we dropped the preregistered inclusion criterion. These findings suggest that the training effect is likely small, but not negligible.

We discuss the largely null results from the current study in the broader literature in two ways. First, the experimental approach to comparing training conditions in number word learning has largely yielded non-significant effects in the literature. In two of the published studies, there was no main effect of training condition on children's performance on the Give-a-Number Task (Gibson et al., 2019; O'Rear & McNeil, 2019; Tillman et al., 2018). These null findings for a main effect of number word training suggest that effects of training may be best interpreted in terms of children's ability to respond to training. For instance, it is possible that the children who benefit most from training are those who already have partial knowledge of the trained number or the counting principles (Bale & Barner, 2009; O'Rear, et al., 2020). Other published work on this topic examines mediators when analyzing the effect of training on children's number knowledge, including children's number gestures and set labelling abilities. Our study adds to this growing body of number word training studies and echoes a need to examine individual differences. Second, despite inconclusive statistical evidence for differences between training conditions on the primary outcome (the 2AFC Task) or the exploratory outcome (pre vs. post-training Give-a-Number Task), our data show that it is possible to train children on the meaning of the next number through a brief exposure that only lasted a few minutes. We found that half of the sample in the experimental group had improved their knower-level, and this is at a rate that is higher than other published work on this topic (e.g., see Gibson et al., 2019; Tillman et al., 2018). This is also at a rate that is higher than expected if children were simply tested twice on the Give-a-Number Task. That is, knower-level differences could reflect measurement error of the task, but we found evidence against this (see Cheung et al., 2024, preprint, for details).Thus, compared to previous training studies that had multiple sessions with manipulatives that had limited success in improving children's number knowledge, our study shows that a simple picture-based comparison could be a fruitful avenue for further research. Increasing the intensity of the training or providing feedback on check trials may increase the effect size.

One unexpected finding from the current study was the poor performance on the training check trials. The current study was adapted from Huang et al. (2011)'s study, who reported that most children passed the training check trials. The critical difference, however, is that our study embeds two comparisons within a trial, rather than one. That is, in our study design, on each trial, children were show, for example, two trained number sets, and then a different number set (e.g., three yellow stars to three red hearts, and then to four yellow stars). This trial structure is likely more taxing for children's attention than showing children a contrast between the trained number set and a different number set ("this card has three birds" "this card does not have three birds", Huang et al. 2011). Our motivation for including this training check trials was based on an interest in testing whether children could *generalize* from training. Success on novel trials was expected if children were able to learn from training trials. However, as pointed out by a reviewer, performance on training check trials may serve as a marker of the training, control or experimental, being successful. We did observe that children were more likely to pass the training check trials in the experimental condition (66%) than in the control condition (57%). We did a post-hoc comparison and found that this difference was not significant ($t(63)$ = 1.4, p = .17). We also recently replicated this pattern of poor performance on check trials in a group of Singaporean preschoolers (N = 31) who completed training similar to the experimental condition in the current study but with fewer training trials (*Mean* on training check = 55%). These results suggest that future studies may need to modify the tripartite trial structure

to improve children's attention and learning from the training stimuli.

Our study is in line with the broader literature on analogical reasoning. We argue that number word learning may be facilitated if we draw on mechanisms that children use to learn other abstract relational concepts (Gentner, 1983; Gentner & Markman 1997; Gentner, 2010). Learning of words that refer to other abstract relational concepts such as 'triangle', 'two-thirds' or 'increasing' are supported if examples enable structural alignment (Christie & Gentner, 2010; Ham & Gunderson, 2019; Smith et al, 2014). Structural alignment is a mechanism of cognitive inference that is enabled if examples fulfill two functions: 1) Highlighting a common relational structure and 2) Highlighting alignable differences (Christie & Gentner, 2010; Gentner, 2010; Gentner & Asmuth, 2019; Gentner & Christie, 2010; Kotovsky & Gentner, 1996). These two functions may be fulfilled when examples allow children to both compare and contrast the meaning of a to-be-learned word (Waxman & Klibanoff, 2000). To highlight a common relational structure, comparisons should vary across any properties irrelevant to the word; to highlight alignable differences, contrasts should only vary in properties relevant to the word (e.g. Namy & Gentner, 1999; Namy & Clepper, 2010; Smith et al., 2014). In the case of number word learning, comparing a set of three red chairs with a set of three blue pencils highlights the common relational structure, namely that both are sets of three, irrespective of different properties of the individual objectsin the sets. Additionally, contrasting three red chairs with four red chairs highlights alignable differences, namely that despite having identical objects, they are not both sets of three. Children may have been limited in their learning of number words in previous studies, because examples did not provide both such comparisons and contrasts and therefore did not enable inference via structural alignment. Although we did not find strong support for our training condition, we speculate that structural alignment likely explains the mechanism underlying number word learning in our study. This tentatively suggests that domain-general learning may play a role in early learning about number word meanings.

We began the study with the assumption that learning the meaning of number word as denoting property of a set, namely the set size, explains why number word learning is hard. The weak effect observed in the current study (a significant effect in one of two analyses) suggests that there are likely other factors that explain children's difficulty in learning number word meanings. Understanding *why* number words are hard to grasp is an important question for further research. How much of number word learning depends on age-related conceptual development, the amount of number input, and learners' ability to attend to or evaluate relevant information for number word meaning? These factors are not mutually exclusive, and it is likely that multiple factors are at play and complement each other in the learning process.

References

Barner, D. (2017). Language, procedures, and the non-perceptual origin of number word meanings. *Journal of Child Language*, *44*(3), 553–590. https://doi.org/10.1017/S0305000917000058

Bloom, P., & Wynn, K. (1997). Linguistic cues in the acquisition of number words. *Journal of Child Language*, *24*(3), 511–533. https://doi.org/10.1017/S0305000997003188

Carey, S., & Barner, D. (2019). Ontogenetic Origins of Human Integer Representations. *Trends in Cognitive Sciences*, *23*(10), 823–835. https://doi.org/10.1016/j.tics.2019.07.004

Carey, S. & Journal of Philosophy, Inc. (2009). Where Our Number Concepts Come From: *Journal of Philosophy*, *106*(4), 220–254. https://doi.org/10.5840/jphil2009106418

Carey, S., Shusterman, A., Haward, P., & Distefano, R. (2017). Do analog number representations underlie the meanings of young children's verbal numerals? *Cognition*, *168*, 243–255. https://doi.org/10.1016/j.cognition.2017.06.022

Chan, J. Y.-C., & Mazzocco, M. M. M. (2017). Competing features influence children's attention to number. *Journal of Experimental Child Psychology*, *156*, 62–81. https://doi.org/10.1016/j.jecp.2016.11.008

Cheung, P., Merkley, R, Wege, T., Jasmin, S., & Ansari, D. (2024). Cross-task variability in assessing number word knowledge. OSF Preprint.

Christie, S., & Gentner, D. (2010). Where Hypotheses Come From: Learning New Relations by Structural Alignment. *Journal of Cognition and Development*, *11*(3), 356–373. https://doi.org/10.1080/15248371003700015

Christie, S., Gentner, D., Call, J., & Haun, D. B. M. (2016). Sensitivity to Relational Similarity and Object Similarity in Apes and Children. *Current Biology*, *26*(4), 531–535. https://doi.org/10.1016/j.cub.2015.12.054

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Gentner, D. (2010). Bootstrapping the Mind: Analogical Processes and Symbol Systems. *Cognitive Science*, *34*(5), 752–775. https://doi.org/10.1111/j.1551-6709.2010.01114.x

Gentner, D., & Asmuth, J. (2019). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, *34*(10), 1298–1307. https://doi.org/10.1080/23273798.2017.1410560

Gentner, D., & Christie, S. (2010). Mutual bootstrapping between language and analogical processing. *Language and Cognition*, *2*(2), 261–283. https://doi.org/10.1515/langcog.2010.011

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56. https://doi.org/10.1037/0003-066X.52.1.45

Gentner, D., & Namy, L. L. (1999). Comparison in the Development of Categories. *Cognitive Development*,

14(4), 487–513. https://doi.org/10.1016/S0885-2014(99)00016-7

Gibson, D. J., Gunderson, E. A., & Levine, S. C. (2020). Causal Effects of Parent Number Talk on Preschoolers' Number Knowledge. *Child Development*, *91*(6). https://doi.org/10.1111/cdev.13423

Gibson, D. J., Gunderson, E. A., Spaepen, E., Levine, S. C., & Goldin-Meadow, S. (2019). Number gestures predict learning of number words. *Developmental Science*, *22*(3). https://doi.org/10.1111/desc.12791

Ham, L., & Gunderson, E. A. (2019). Utilizing analogical reasoning to aid children's proportional reasoning understanding. *Journal of Numerical Cognition*, *5*(2), 140–157. https://doi.org/10.5964/jnc.v5i2.193

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. https://doi.org/10.1177/0956797610363552

Kotovsky, L., & Gentner, D. (1996). Comparison and Categorization in the Development of Relational Similarity. *Child Development*, *67*(6), 2797. https://doi.org/10.2307/1131753

Marchand, E., & Barner, D. (2018). Analogical Mapping in Numerical Development. In *Language and Culture in Mathematical Cognition* (pp. 31–47). Elsevier. https://doi.org/10.1016/B978-0-12-812574-8.00002-X

Matlen, B. J., Gentner, D., & Franconeri, S. L. (2020). Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(5), 443–457. https://doi.org/10.1037/xhp0000726

Mazzocco, M. M. M., Chan, J. Y.-C., Bye, J. K., Padrutt, E. R., Praus-Singh, T., Lukowski, S., Brown, E., & Olson, R. E. (2020). Attention to numerosity varies across individuals and task contexts. *Mathematical Thinking and Learning*, *22*(4), 258–280. https://doi.org/10.1080/10986065.2020.1818467

Merkley, R., Scerif, G., & Ansari, D. (2017). What is the precise role of cognitive control in the development of a sense of number? *Behavioral and Brain Sciences*, *40*, e179. https://doi.org/10.1017/S0140525X1600217X

Mix, K. S. (1999). Similarity and Numerical Equivalence. *Cognitive Development*, *14*(2), 269–297. https://doi.org/10.1016/S0885-2014(99)00005-2

Mix, K. S., & Sandhofer, C. M. (2007). Do we need a number sense?. In *Integrating the Mind: Domain General Versus Domain Specific Processes in Higher Cognition*, 293. https://doi.org/10.4324/9780203926697

Mix, K. S. (2008). Children's equivalence judgments: Crossmapping effects. *Cognitive Development*, *23*(1), 191–203. https://doi.org/10.1016/j.cogdev.2007.03.001

Mix, K. S., Sandhofer, C. M., Moore, J. A., & Russell, C. (2012). Acquisition of the cardinal word principle: The role of input. *Early Childhood Research Quarterly*, *27*(2), 274–283. https://doi.org/10.1016/j.ecresq.2011.10.003

Namy, L. L., & Clepper, L. E. (2010). The differing roles of comparison and contrast in children's categorization. *Journal of Experimental Child Psychology*, *107*(3), 291–305. https://doi.org/10.1016/j.jecp.2010.05.013

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). https://doi.org/10.1111/desc.12819

O'Rear, C. D., McNeil, N. M., & Kirkland, P. K. (2020). Partial knowledge in the development of number word understanding. *Developmentalscience*, 23(5). https://doi.org/10.1111/desc.12944

Posid, T., & Cordes, S. (2018). How high can you count? Probing the limits of children's counting. *Developmental Psychology*, *54*(5), 875–889. https://doi.org/10.1037/dev0000469

Slusser, E. B., & Sarnecka, B. W. (2011). Find the picture of eight turtles: A link between children's counting and their knowledge of number word semantics. *Journal of Experimental Child Psychology*, *110*(1), 38–51. https://doi.org/10.1016/j.jecp.2011.03.006

Slusser, E., Stoop, T., Lo, A., & Shusterman, A. (2017). Children's use of newly acquired number words in novel contexts. In *Poster presented at the Biennial Meeting for the Society for Research in Child Development, Austin, TX*.

Smith, L., Ping, R. M., Matlen, B. J., Goldwater, M. B., Gentner, D., & Levine, S. (2014). Mechanisms of Spatial Learning: Teaching Children Geometric Categories. In C. Freksa, B. Nebel, M. Hegarty, & T. Barkowsky (Eds.), *Spatial Cognition IX* (Vol. 8684, pp. 325–337). Springer International Publishing. https://doi.org/10.1007/978-3-319-11215-2_23

Sullivan, J., & Barner, D. (2011). Number words, quantifiers, and principles of word learning: Number words, quantifiers, and principles of word learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 639–645. https://doi.org/10.1002/wcs.140

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*, *36*(5), 571–581. https://doi.org/10.1037/0012-1649.36.5.571

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193. https://doi.org/10.1016/0010-0277(90)90003-3

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, *24*(2), 220–251. https://doi.org/10.1016/0010-0285(92)90008-P