Cognition 200 (2020) 104243

Contents lists available at ScienceDirect

Cognition

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

Learning the generative principles of a symbol system from limited examples

Lei Yuan[a,*], Violet Xiang[b], David Crandall[b], Linda Smith[c,d]

[a] Department of Psychological and Brain Sciences, Indiana University, United States of America
[b] School of Informatics, Computing, and Engineering, Indiana University, United States of America
[c] Department of Psychological and Brain Sciences, Indiana University, United States of America
[d] School of Psychology, University of East Anglia, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:
Associative learning
Statistical learning
Deep learning
Generative learning
Symbol systems
Education

ABSTRACT

The processes and mechanisms of human learning are central to inquiries in a number of fields including psychology, cognitive science, development, education, and artificial intelligence. Arguments, debates, and controversies linger over the questions of human learning with one of the most contentious being whether simple associative processes could explain human children's prodigious learning, and in doing so, could lead to artificial intelligence that parallels human learning. One phenomenon at the center of these debates concerns a form of far generalization, sometimes referred to as "generative learning", because the learner's behavior seems to reflect more than co-occurrences among specifically experienced instances and to be based on principles through which new instances may be *generated*. In two experimental studies (N = 148) of preschool children's learning of how multi-digit number names map to their written forms and in a computational modeling experiment using a deep learning neural network, we show that data sets with a suite of inter-correlated imperfect predictive components yield far and systematic generalizations that accord with generative principles and do so despite limited examples and exceptions in the training data. Implications for human cognition, cognitive development, education, and machine learning are discussed.

1. Introduction

There are two different stories that one can tell about human learning (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McClelland et al., 2010). In some tasks, learning is slow with generalization requiring extensive experience with many examples, and even then, generalization may be limited and error-prone (Bion, Borovsky, & Fernald, 2013; Fuson & Briars, 1990; Gentner, 2010; McMurray, Horst, & Samuelson, 2012). Many categories of school learning including early reading and mathematics seem to fit this description (Chi, Kristensen, & Roscoe, 2012; Siegler & Lortie-Forgues, 2017). However, in other contexts, human learning appears much less data-hungry and can be characterized as showing extensive generalization from limited experience with a small portion of possible instances (Aslin, 2017; Carey & Bartlett, 1978; Casler & Kelemen, 2005). Generalization from a few examples is sometimes known as "few-shot learning" and has been documented in domains such as object recognition (Krizhevsky, Sutskever, & Hinton, 2012), letter recognition (Lake, Salakhutdinov, & Tenenbaum, 2015), and word learning by children (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; F. Xu & Tenenbaum, 2007). For example, typically-developing 2.5-year-old children appropriately extend a newly-heard object name to new instances of the category given experience with just one named object from that category (Landau, Smith, & Jones, 1988; Smith et al., 2002; Smith, Jones, & Landau, 1996).

Rapid and far generalization has also been characterized as a form of "generative learning" because the learner seems not to just learn about specifically experienced instances but rather to learn principles through which new instances may be *generated* (Lake, Linzen, & Baroni, 2019; Son, Smith, & Goldstone, 2012). For example, typically-developing preschool children learning English can generate the regular plural form for a seemingly unlimited number of nouns, needing only one exposure to the singular form of the noun to do so (Berko, 1958; Brown, 1973; Mervis & Johnson, 1991; Treiman, 1993). Given that human learning often seems slowly incremental and limited in generalizability, these cases of principled far generalization have attracted considerable research attention in domains as diverse as cognitive development and machine learning (Fe-Fei, Fergus, & Perona, 2003; Imai, Gentner, & Uchida, 1994; Kemp, Perfors, & Tenenbaum, 2007; Lake et al., 2019; Smith & Samuelson, 2006).

Current theoretical debates are focused on the learning mechanisms. By most accounts, the critical factor is *prior* knowledge of the principles

* Corresponding author.
E-mail address: leiyuan@indiana.edu (L. Yuan).

https://doi.org/10.1016/j.cognition.2020.104243
Received 29 November 2017; Received in revised form 19 February 2020; Accepted 20 February 2020

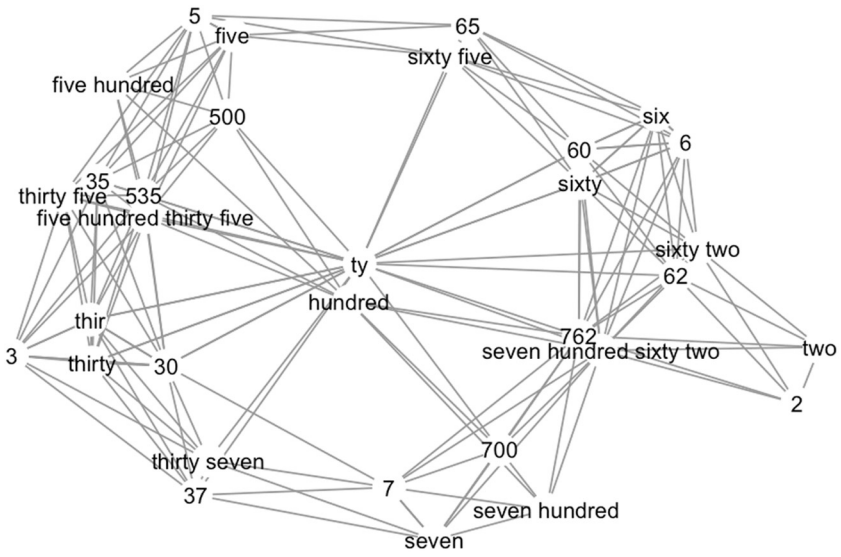

Fig. 1. An illustration of all possible partial mappings within and between written number symbols and spoken number names for four randomly chosen numbers 37, 65, 535 and 762. The nodes depict individual components of written number symbols or spoken number names. The edges depict co-occurrences and partial mappings among the nodes. As can be seen, there are massive overlapping and redundant connections among pairs of written symbols and their component names that instantiate the to-be-learned generative principles.

for representing instances within the to-be-learned domain (Fe-Fei et al., 2003; Griffiths et al., 2010; Lake et al., 2015; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). These prior principles could be domain-specific and part of human core (and innate) knowledge systems (Gopnik & Bonawitz, 2015; Spelke, 2016), making rapid generalization a specialization for only some core domains. Others have suggested that these generative principles can be discovered through more general learning mechanisms and in many different domains. By some accounts, associative learning mechanisms may be sufficient (e.g., Botvinick & Plaut, 2004; Colunga & Smith, 2005; Elman, 1990; McClelland et al., 2010; Rogers & McClelland, 2004) but others have argued that associative mechanisms—even in their most currently advanced forms such as deep learning neural networks—are fundamentally limited, requiring extensive training and even then can only approximate the learning of generative principles (Griffiths et al., 2010; Tenenbaum et al., 2011). By these accounts, more powerful statistical learning mechanisms and computations are required to discover generative principles from limited training data (Kemp, Goodman, & Tenenbaum, 2008; Kemp, Perfors, & Tenenbaum, 2007; Rule, Dechter, & Tenenbaum, 2015; F. Xu & Tenenbaum, 2007).

In these debates and the related experimental and theoretical studies, there has been little consideration of the properties of data structures that support principled and far generalization from limited training experiences. However, debates about learning mechanisms cannot be divorced from the data structures on which those mechanisms operate as all learning depends on the learning mechanisms, the statistical structure of the experiences on which they operate, and the match between the two (Dupoux, 2018). The real-world cases (e.g., novel object name generalization, regular plural forms of English nouns) used to document human propensity for principled far generalization have data structures different from those used in most laboratory studies (but see, Billman, 1989, 1993; Billman & Knutson, 1996). They are characterized by multiple inter-predictive features that are redundant, degenerate, overlapping, imperfect, and that offer *multiple* pathways to generalization (Bloom et al., 2006; Colunga & Smith, 2005; MacWhinney, Leinbach, Taraban, & McDonald, 1989; Yoshida & Smith, 2003). By hypothesis, a suite of inter-correlated imperfect predictive components can give rise to generalizations that accord with generative principles and can do so despite limited training data, exceptions, and idiosyncratic individual experiences.

Here we provide initial evidence for this hypothesis by showing that preschool children and a general-purpose deep-learning neural network trained on a limited data set with a multiple inter-predictive structure show principled extensive generalization. The domain we use to make this initial case is the human-invented symbol system through which we name and write multi-digit numbers. We chose this domain for five reasons. First, it is a real-world system but, as a relatively recent human invention, it is a knowledge domain without specifically evolved core mechanisms. Second, it is well documented that the base-10 notational system—and the multiplicative hierarchical structures that underlie it—are difficult for school-aged children to master (Fuson, 1988; Mann, Moeller, Pixner, Kaufmann, & Nuerk, 2012; Ross, 1995). Third, there is suggestive evidence that at least some preschool children know how to map never-before-encountered multi-digit number names to their written forms, despite likely minimal experience with the names and written forms of multi-digit numbers (Mix, Prather, Smith, & Stockton, 2014; Yuan, Prather, Mix, & Smith, 2019). Fourth, and as we expand below, spoken and written number names have a data structure of co-predicting *surface* features that are redundant, overlapping, and imperfect, but provide multiple paths to correct generalization. Fifth, this case provides a grounding for consideration of the distinction (and potential relation) between generalization that is *consistent* with generative principles versus the explicit *representation* of those principles (Lake, Ullman, Tenenbaum, & Gershman, 2016; Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015).

Informal experiences of hearing the spoken names for written multi-digit numbers and seeing their corresponding written forms—for example hearing “seven hundred sixty-two” while seeing “762”—comprise a data set of potential interest for learning about place value (Grossberg & Repin, 2003; Rule et al., 2015). Fig. 1 illustrates—for a very small set of possible numbers—the many redundant and overlapping mappings (represented by the edges) *among the surface structure* of written numbers and their spoken names (represented by the nodes). For example, in the written form “535” there are two “5”s, one on the far left and one on the far right, and in the spoken name “five hundred and thirty-five,” “five” occurs twice, in the first position and in the last. The written form “3” systematically co-occurs with “three,” “thirty,” and “three hundred.” “Thirty” and “sixty” both end in “-ty,” and in their co-occurring written forms, the digits named with a “-ty” appear just before the last (rightmost) position in the string of digits. “Eighty” and “ninety” (but not “eleven” nor “twenty”) contain the name most strongly associated with the written form of a single digit (8 and 9). These patterns provide a hodgepodge of paths to mapping a heard multi-digit name to its written form. As we consider in the General Discussion, they also may provide a path to a deeper understanding of the generative principles that are the source of these exploitable surface properties.

Both spoken and written forms have their origins in underlying

principles of the base-10 multiplicative hierarchy of places. Thus, “762” and “seven hundred and sixty-two” each refer to the same decomposition of the quantity: to 7 sets of 100, 6 sets of 10, and 2 sets of 1 with 100 equal to 10 sets of 10, and 10 equal to 10 sets of 1. Ultimately, children need to explicitly understand these principles if they are to successfully calculate with multi-digit numbers. But, by hypothesis, they do not need knowledge of the underlying multiplicative hierarchy to map any heard number to its written form; all they need to do is exploit the plethora of predictive surface properties to map number names to written forms. These multiple predictive surface properties linking names to written forms will not lead to perfect performance (because they are imperfect and local predictors); but, by hypothesis, they can lead to far generalizations at levels well above chance in mapping newly encountered individual multi-digit number names to their written forms.

In contrast to this characterization of possible early knowledge of multi-digit numbers, the consensus view on the development of place value concepts is that the mapping of number names to written multi-digit numbers is hard and error-filled even for school age children and tightly tied to understanding the underlying base-10 principles (Fuson & Kwon, 1991, 1992; Geary, Bow-Thomas, Liu, & Siegler, 1996; Ho & Fuson, 1998). For the most part, this conclusion derives from studies of school-age children's understanding of the underlying base-10 principles, studies that find predictive errors in naming written forms and in calculating with multi-digit numbers (Cooper & Tomayko, 2011; Fuson & Kwon, 1991) and studies focused on children's difficulties with the exceptions in the naming system (e.g., the teens, Miura & Okamoto, 1989; Saxton & Towse, 1998). The general conclusion is that explicit formal training of the notational principles is essential to both understanding the notational system and to using it to calculate (Fuson, 1986; Fuson & Briars, 1990). From these findings, the general view in the education literature and education practice is that introducing multi-digit numbers is best delayed until the start of formal teaching about the base-10 system, typically first or second grade (Fuson, 1986; Hanich, Jordan, Kaplan, & Dick, 2001; Kamii, 1986).

However, several recent studies indicate that at least some preschool children know how number names map to written digits, performing well above chance when asked to pick the written version of three- and four-digit numbers given the spoken name (Byrge, Smith, & Mix, 2014; Mix et al., 2014; Yuan et al., 2019), for example, choosing 836 over 834 or 863, given the spoken name of “eight-hundred and thirty-six.” Less clear is how these children learned whatever knowledge allowed them to succeed in this task. Considerable evidence indicates that number talk to preschool children is quite sparse and talk about multi-digit numbers is exceedingly rare (Dehaene, 1992; Dehaene & Mehler, 1992; Levine, Suriyakham, Rowe, Huttenlocher, & Gunderson, 2010). By these estimates, then, the likelihood that the preschool children showing early competence in mapping names to multidigit numbers had encountered the name and written form of any particular 3-digit number (e.g., 836) tested in these previous studies is vanishingly small. We propose that the children who performed well acquired the *general* ability to map heard names to written multi-digit numbers from limited exposure through learning mechanisms that exploit the multiple correlated—albeit imperfect—regularities that link number names and written forms.

We test this hypothesis in three studies. The first two are experimental studies that show that preschool children show systematic generalization in mapping the names to written forms given minimal exposure to a small set of multi-digit numbers and their names. The third study is a computational modeling experiment. The purpose of this modeling experiment is not to provide a complete or accurate model of children's internal learning mechanisms but rather to show that an associative learning mechanism given a data set with imperfect, redundant local predictors will exhibit far generalization. To this end, we used a general-purpose deep neural network trained similarly to the children in the two experiments. The modeling experiment provides evidence for generalization *consistent* with generative principles without explicit representation of those principles.

2. Study 1

2.1. Participants

The final sample consisted of forty preschool children (mean age: 4.5 years, range: 3.16–5.94 years) from a Midwestern town in the United States. There were 18 females and 22 males. Families were contacted about the study through a consented database or through local preschools and day care centers that served families from a wide range of economic circumstances. Informed consent was obtained from each participant's legal guardian prior to the study. Each child participated in five successive sessions (pre-test, 3 days of training, post-test) on separate days of the week (i.e., Monday to Friday). If the child missed one and only one session during the week, he or she participated on the next available weekday. Five additional participants were excluded from the study due to missing one or more sessions during the study. Forty children participated in the training condition; an additional seventeen children participated in a no-training control condition included to check on test-retest effects. On pre-test and post-test days (but not training days), some children also participated in other tasks (including magnitude judgements) that were components of other experiments being conducted in these same schools and daycares.

2.2. Stimuli and procedure

2.2.1. Training

The training was designed to present children with minimal training and minimal experience with specific multi-digit numbers. The selection of training numbers was designed to mimic likely real-world experiences of young children in which a few single- and double-digit numbers were repeated with most 3- to 4-digit numbers encountered only once (Dehaene, 1992; Levine et al., 2010). There were 18 trials on each of the 3 training days for a total of 54 learning trials for the entire study. Across the 3 days of training and total of 54 trials, children heard the names and saw the written forms of 36 unique numbers that varied from 1- to 4-digit numbers. Of the 36 unique numbers, 12 were repeated during training and each of the 24 other unique instances occurred *just once* in training. Three-digit numbers were named with the word “hundred” as in “three-hundred fifty-two” and 4-digit numbers were pronounced with the word “thousand” as in “two-thousand five-hundred twenty-one”.

Training was embedded in casual learning activities meant to mimic possible everyday contexts through which preschool children might encounter multi-digit numbers and their names. The contexts were designed so that there was no explicit teaching or mention of the underlying syntactic rules and no specific task with strictly defined right or wrong responses from the children. Rather, children were simply encouraged to follow along and have fun with two engaging activities: storybook reading and making numbers with cards. We used two training orders, one in which similar numbers (e.g., 223, 224) occurred in close proximity (N = 20) and one in which the order was randomly determined (N = 20). These different orders had no effects that approached statistical significance (see supplementary material) and are not considered further. Table 1 shows all training numbers and Fig. 2 shows the training materials.

For **the Storybook reading** component of the training, three picture books (one for each training session) were created with each containing four stories. Each page was printed on A4-sized horizontally arranged paper. Most of the pages consisted of a cartoon caricature (roughly 2 in. tall and 1 in. wide), some objects (roughly 2 in. tall and 1 in. wide) and printed multi-digit numbers (see Fig. 2 for an example). The numbers were printed in 42-point Arial font. Each story had five pages which were put into a clear sheet protector and stored in a

Table 1
All training activities and numbers used in Study 1.

Activity	Order	Trial	Training day 1	Training day 2	Training day 3	Total subjects
Storybook	Grouped	1	2	2	14	20
Storybook	Grouped	2	3	3	15	20
Storybook	Grouped	3	4	4	16	20
Storybook	Grouped	4	223	125	515	20
Storybook	Grouped	5	224	135	525	20
Storybook	Grouped	6	225	145	535	20
Storybook	Grouped	7	40	14	2	20
Storybook	Grouped	8	60	15	3	20
Storybook	Grouped	9	70	16	4	20
Storybook	Grouped	10	402	250	2520	20
Storybook	Grouped	11	502	350	3520	20
Storybook	Grouped	12	602	450	4520	20
Make-a-number	Grouped	13	14	1000	21	20
Make-a-number	Grouped	14	15	2000	121	20
Make-a-number	Grouped	15	16	3000	221	20
Make-a-number	Grouped	16	470	21	40	20
Make-a-number	Grouped	17	570	121	60	20
Make-a-number	Grouped	18	670	221	70	20
Storybook	Random	1	502	3	525	20
Storybook	Random	2	2	135	3	20
Storybook	Random	3	60	250	4520	20
Storybook	Random	4	14	16	21	20
Storybook	Random	5	402	1000	16	20
Storybook	Random	6	670	21	40	20
Storybook	Random	7	224	2	515	20
Storybook	Random	8	570	145	2	20
Storybook	Random	9	602	450	3520	20
Storybook	Random	10	470	15	121	20
Storybook	Random	11	70	3000	15	20
Storybook	Random	12	3	221	60	20
Make-a-number	Random	13	40	4	535	20
Make-a-number	Random	14	15	125	4	20
Make-a-number	Random	15	225	350	2520	20
Make-a-number	Random	16	4	14	221	20
Make-a-number	Random	17	223	2000	14	20
Make-a-number	Random	18	16	121	70	20

binder. A sample story about saving money is illustrated on Fig. 2. The experimenter first presented Page A and explained to the child, “Johnny wants to save money to buy his favorite food and toys. Do you want to see what Johnny wants to buy?” Then she presented Page B and said, “He wants to buy a big cake.” Pointing to the written number, the experimenter asked, “Do you know how much it costs?” Children were not expected to and typically did not respond to this rhetorical question but regardless of the nature of any response, the experimenter immediately said, “It costs forty (pointing to the digits sequence, i.e.,“4” followed by “0”) dollars.” She then repeated the number once more, still pointing to the written digits in sequence while saying, “The big cake costs forty dollars.” She next presented Pages C and D in an identically structured narrative with the only change being the object's name and the corresponding numbers. On Page E, the experimenter asked the child, “Can you tell me which thing costs the most money?” Regardless of the child's response (or nonresponse), the experimenter immediately stated the relation—“The cake costs forty dollars (while pointing to the digits in the written numeral “40” in sequence); the bicycle costs sixty dollars (pointing to the digits in the written numeral “60” in sequence); and the toy car costs seventy dollars (pointing to the digits in the written numeral “70” in sequence). So, the item that costs the most money is the car (point to the toy car).” This mention of relative magnitudes was included to encourage children to connect and *compare* the number names and written forms for different quantities which has been shown to highlight the common relational structures (Gentner, 1983; Gentner et al., 2016; Kotovsky & Gentner, 1996; Yuan, Uttal, & Gentner, 2017), and in this case the many predictive elements characterizing multi-digit number names and their written forms.

In the **Make-a-number game**, two identical sets of number cards were created to be used by the experimenter and the participant. The cards were made from 1-inch by 2-inch foam sheet and number stickers. Each card depicted just one digit that had the dimension of roughly 1 (width) by 2 (height) inch. During the training, the experimenter first made a number using her set of digit cards. For example, she first told the child what number they were going to make: “We are going to make two hundred thirty-five. Watch me, I am going to make two hundred thirty-five.” She then picked up the card “2” and said, “I need a 2 for two hundred” while putting it down on the table. She then picked up the card “3” and said, “I need a 3 for thirty” while putting it on the right of the card “2”. Lastly, she picked up the card “5” and said, “I need a 5 for five” while putting it on the right of the card “3”. She then invited the child to make the same number by saying “Can you make two hundred thirty-five?” This task only required the child to copy the just-preceding behavior of the experimenter and the still-in-view example. If the child had trouble doing so, the experimenter coached the child to make the correct number in a naturalist way, such as reminding the child that “We need a 2 for two hundred.” After the child finished making the number, the experimenter asked, “What number did you just make?” Regardless of the child's response, the experimenter repeated the name of the number one more time by saying, “Good job. You just made two hundred thirty-five.” Again, the goal of training was only to expose children to corresponding names and numbers and in an engaging and active way.

2.2.2. Pre- and posttests

The Which-is-N test (Mix et al., 2014; Yuan et al., 2019) is a commonly-used measure of children's ability to map spoken names to written numbers. The structure of the task, a two-alternative forced choice between two written forms given a spoken name, differs from the casual structure of storybook reading and the make-a-number game. There were 16 test items: 8-vs-2, 15-vs-5, 12-vs-22, 11-vs-24, 85-vs-850, 105-vs-125, 201-vs-21, 206-vs-260, 36-vs-306, 350-vs-305, 402-vs-42, 64-vs-604, 670-vs-67, 807-vs-78, 1000-vs-100, 1002-vs-1020. *All* test pairs included at least one number *never* seen in training. For half the test items, one (but not both) of the choice numbers (but not necessarily the target) was presented during training; for the remining test items, both choice numbers were novel. In this way, the test is a strong measure of generalization. Single digit numbers were included to provide children with some easy trials and avoid floor effects. The choice items were presented on an A4-sized page in a binder. The numbers were printed in 42-point Arial font and were arranged horizontally across the center of the page. Two sets of orders (Set A, Set B) were created and counterbalanced across subjects.

Each child completed five sessions (pre-test, three training sessions, post-test), and each lasted 10 to 18 min, appropriate to the attentional abilities of preschool children. 30% of participants (we substantially increased this proportion in Study 2) were blind tested by an experimenter who was not aware of the conditional assignment of the participant. There was no significant difference in the learning outcome between children who were blind tested and those who were tested and trained by the same experimenter (see supplemental materials).

2.2.3. Baseline measures of improvement

Although unlikely, children could, in principle, show improved performance at post-test because of a test-retest effect or because of increasing comfort with experimenters. Accordingly, an additional group of children (n = 17) participated in an identical training to the main experimental condition, but instead of spoken number names and written forms, their training involved spoken words and their written forms. For example, in the storybook reading activity shown in Fig. 2, the experimenter said, “Johnny wants to buy a big cake.” She then pointed to the letter “C,” and asked, “What letter is this?” Regardless of the child's response, the experimenter would say, “It is C. C for cake. Johnny wants to buy a big cake.” Later, the experimenter asked, “So what does Johnny want to buy? C (point to the letter C) for cake, B

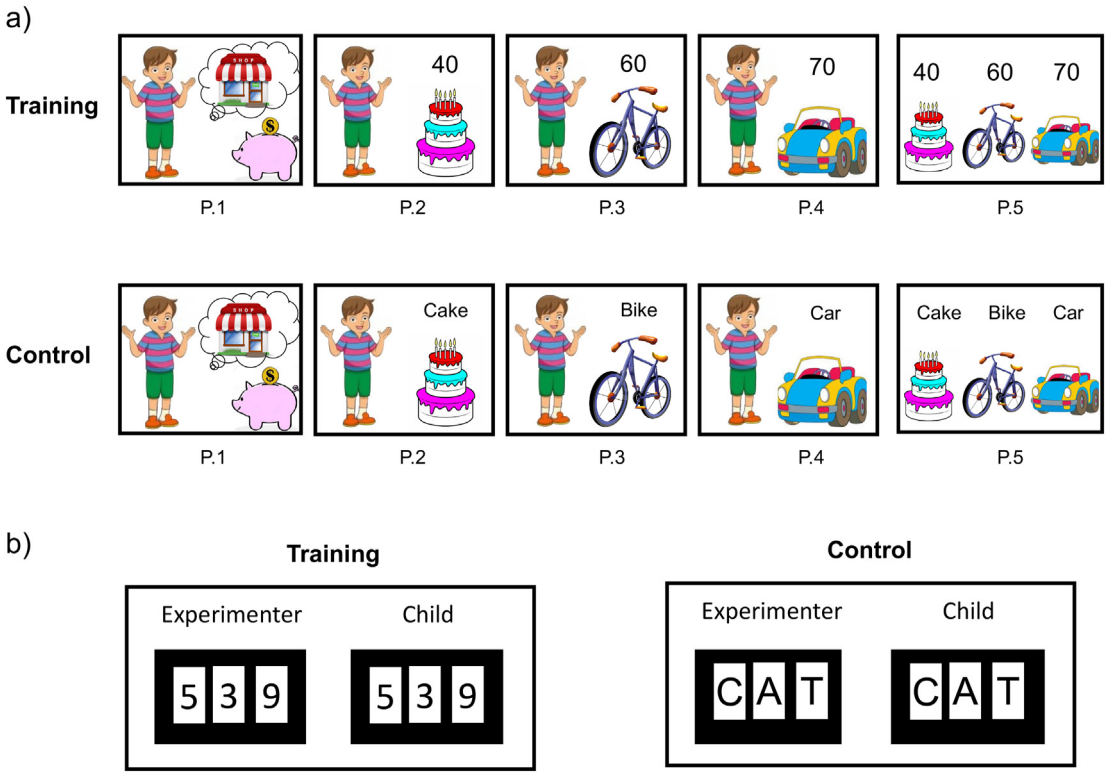

Fig. 2. a) Sample book from the storybook reading activity in Study 1. Each book has 5 pages from P.1 to P.5. Materials were identical between the training condition (top row) and the control condition (bottom row), with the only difference being that the training condition involved numbers, and the control condition involved spelling. b) Sample materials for the making numbers (or words) with cards game in Study 1. The child and the experimenter each have a set of cards; the only difference between the training and the control condition is the content on the cards—individual digits for the training and individual letters for the control condition.

(point to the letter B) for bike, and C (point to the letter C) for car." The Make-a-word game was the same as the Make-a-number game except that the child and experimenter spelled words using letter cards. For a sample of the words used, see Supplemental Materials. Children were tested in the same number pretest and post-tests tasks as the children in the main experiment. There was also no significant difference in the ages or pretest scores between participants in the training condition and those in the baseline measures condition (see supplemental materials). Children in the baseline measures condition showed no increase in performance on post-test relative to pretest. A Linear Mixed Effect Model was conducted in which time was entered as a fixed effect and participant was entered as a random effect. The model failed to find a main effect of time, F (1, 16) = 0.08, p = .79. Accuracy at pre-test (M = 0.60, SE = 0.03) and at post-test (M = 0.59, SE = 0.03), t (16) = 0.27, p = .79, d = 0.07, did not differ. Thus, pre- and post-test effects or similar experiences with the experimenter appear at best minimal.

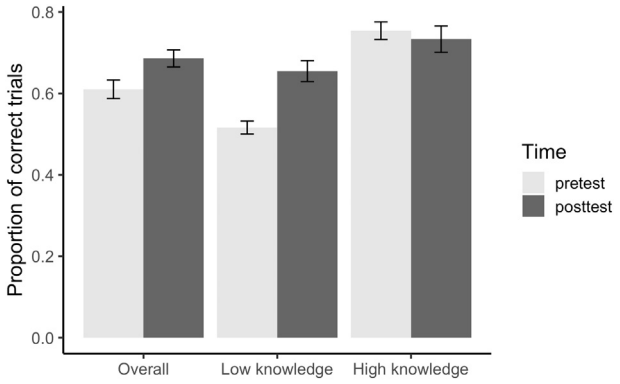

Fig. 3. Proportion of correct trials at pretest and posttest for all children and children with different levels of early knowledge in Study 1 (defined as above or below 65% accuracy at pretest). Error bars indicate standard errors.

2.3. Results and discussion

Children from the main Training Experiment showed modest above chance performance at pretest, t (39) = 4.49, p < .0001, M = 0.61, SE = 0.03, d = 0.71, consistent with previous studies showing that some preschool children have early multi-digit number knowledge (Mix et al., 2014; Yuan et al., 2019). As can be seen in Fig. 3, these children improved significantly from pretest (M = 0.61, SE = 0.02) to posttest (M = 0.69, SE = 0.02), t (39) = 3.69, p < .001, d = 0.58. This training effect was also confirmed by a Linear Mixed Effect model (LMM) conducted in the R environment (Team, 2017) using the lme4 package (Bates et al., 2015). Significance values were obtained using the Afex package (Singmann, Bolker, Westfall, & Aust, 2015) with the KR method, which uses the Kenward-Roger's approximation to calculate the p values (Luke, 2017). Time (pre- or post-test) was entered as a fixed effect and participant was entered as a random effect. There was a significant main effect of time, F (1, 39) = 13.64, p < .001, again indicating an effect of training.

The scientific significance of the improvement, even though its absolute magnitude might seem small (an average increase of 8%), arises from the minimal nature of the training (3 days, a total of 36 unique numbers with just a few repetitions) and was evident on novel test items. A direct test of performance on partially novel and totally novel test pairs, excluding items that may be solved by knowledge of single-digit numbers alone (i.e., "8 vs 2", "15 vs 5"), revealed no significant difference, t (39) = 1.68, p = .10, d = 0.27, in post-test performance on the two classes of test items, consistent with the predicted

far generalization from minimal learning to novel items. This conclusion is also supported by a comparison of pre- and post-test performance on the test items in which both target and foil were novel. A Linear Mixed Effect Model was conducted using only the trials in which both the names and the written choices were novel; time was entered as a fixed effect and participant was entered as a random effect. Results showed a significant main effect of time, F (1, 39) = 7.32, p = .01. For these totally novel test items, children's performance significantly improved from pretest (M = 0.54, SE = 0.04) to posttest (M = 0.63, SE = 0.03), t (39) = 2.71, p = .01, d = 0.43.

To further explore how overall learning was related to individual factors, a multiple linear regression was conducted using age, gender and pretest score to predict learning (defined as changes in scores from pretest to posttest). Learning was not related to gender (b = −0.02, p = .62), but modestly and positively related to continuous age (b = 0.05, p = .03) with pretest score being the most predictable factor in how much children learned from the training (b = −0.55, p < .001). As can be seen in Fig. 3, children with the *lowest* pretest scores increased the most from pre-test to post-test, a finding that also supports the effectiveness of the limited training exposure to number names and written multi-digit numbers.

The results of Experiment 1 provide initial support for the hypothesis that—given the right data structure—minimal experience with a relatively few instances from the entire domain (in the present case, all numbers up to 9999) can lead to broad generalized knowledge to novel instances sampled from that same domain. Moreover, the core of the training was simply exposure to the corresponding spoken names and written forms. There was no special teaching method or explicit explanation of why or how multi-digit numbers work as they do.

3. Study 2

Study 2 tested the robustness of the training effect observed in Study 1 with four modifications. First, children received either the storybook reading or the making numbers with cards training; in this way, the experiment provides evidence for the idea that that exposure to corresponding number names and written forms—not the particular activity—is the key factor in learning. Second, we generated an entirely new set of training numbers to show that the effects of Study 1 were not driven by the specific 36 unique training instances chosen for that study. Third, to provide a more sensitive test of the effects of training, we excluded children who performed above 85% correct on the pretest. Fourth, to provide a more sensitive test of the potential effect of learning specific items, we also counterbalanced whether numbers that appeared during the training were the target or the foil number during testing (which was not done in Study 1).

3.1. Participants

The final sample (66 in the Main Experiment and 25 in the Control measures) were recruited from the same general population as Study 1. The mean age of the participants was 4.4 years (range: 2.89–5.99). There were 49 males and 42 females. The experiment settings and timelines were identical to Experiment 1. There were two training activities (N = 30 and N = 36) as described below. Eight participants were excluded from the study due to missing one or more sessions during the study. An additional fifteen participants were excluded due to pretest scores higher than 85% correct.

3.2. Stimuli and procedures

3.2.1. Main training experiment

As shown in Fig. 4, the procedures in the training conditions were identical to Study 1 with two exceptions. First, a new set of training numbers (shown in Table 2) was selected to follow the same distributional properties (e.g., numbers with 1- to 4-digits numbers, repetitions) as Study 1 but differed in the specific multi-digit numbers used (which were randomly selected from possible numbers fitting the distributional constraints). Second, each participant received only one condition—either the storybook reading activity (N = 36) or the making numbers with cards activity (N = 30). As in Study 1, each participant received 18 training trials on each day with 54 trials in total across three training sessions. **Pre- and post-tests**. The pre- and posttest items and the procedures for administering them were identical to Study 1 with the exception of the counter balanced designation as target or foil of the 8 training items that were the post-test. 64% of participants were blind tested by an experimenter who was not aware of the conditional assignment of the participant. There was no significant difference in the learning outcome between children who were blind tested and those who were tested and trained by the same experimenter (see supplemental materials).

3.2.2. Baseline measures of improvement

Again, to rule out possible test-retest effects and exposure to the experimenter, we again collected pre- and post-test data from children who did not experience corresponding spoken names and written forms of multi-digit numbers. For this pre- to post-test measure, a separate group of children (N = 25) was exposed to just one stream of the training information (i.e., either the auditory or visual stream) that was used in the main training condition (see Fig. 4). These experiences are near identical to the training experiences in the storybook training condition but differ only in missing the training in mapping the heard number name to the written form. The children who received only one stream of information (either written numbers or spoken number names) did not demonstrate learning. A Linear Mixed Effect Model was conducted in which time was entered as a fixed effect and participant was entered as a random effect. Results failed to find a significant main effect of time, F (1, 24) = 1.08, p = .31. Children's performance at pretest (M = 0.55, SE = 0.03) and at posttest (M = 0.59, SE = 0.04) did not differ significantly, t (24) = 1.04, p = .31, d = 0.21. These results again suggest minimal if any test-retest effects, minimal effects of familiarity with the experimenters, or with one modality of the training information but no association between names and numbers. Further, there was no significant difference in the ages or pretest scores between participants in the main training condition and those in the baseline measures condition (see supplemental materials).

3.3. Results and discussion

Consistent with Study 1, children from the training condition showed modest but above chance performance at pretest, t (65) = 4.89, p < .001, d = 0.60, M = 0.58, SE = 0.02. To examine the training effect, a Linear Mixed Effect model was conducted in which time (i.e., pretest or posttest) and training activity (i.e., storybook reading or making numbers with cards) were entered as fixed effects and participant was entered as a random effect. Results showed a significant main effect of time, F (1, 64) = 11.80, p < .001. Neither the effect of training activity, F (1, 64) = 0.52, p = .47, nor the interaction between time and training activity reached significance level, F (1, 64) = 0.16, p = .69. Overall, as can be seen in Fig. 5, children who received training improved significantly from pretest (M = 0.58, SE = 0.02) to posttest (M = 0.65, SE = 0.02), t (65) = 3.5, p < .001, d = 0.43. The lack of differences between the two training formats suggests that the nature of the activity—listening to a story or actively building numbers—is not a key factor. What is similar across the two training activities is exposure to co-occurring multi-digit number names and their written forms.

A direct test of performance on partially novel and totally novel test items, excluding items that may be solved by knowledge of single-digit numbers alone (i.e., “8 vs 2”, “15 vs 5”), revealed no significance difference between these two classes of test items, t (65) = 1.18, p = .24, d = 0.15. Focusing only on the trials in which both target and foil were

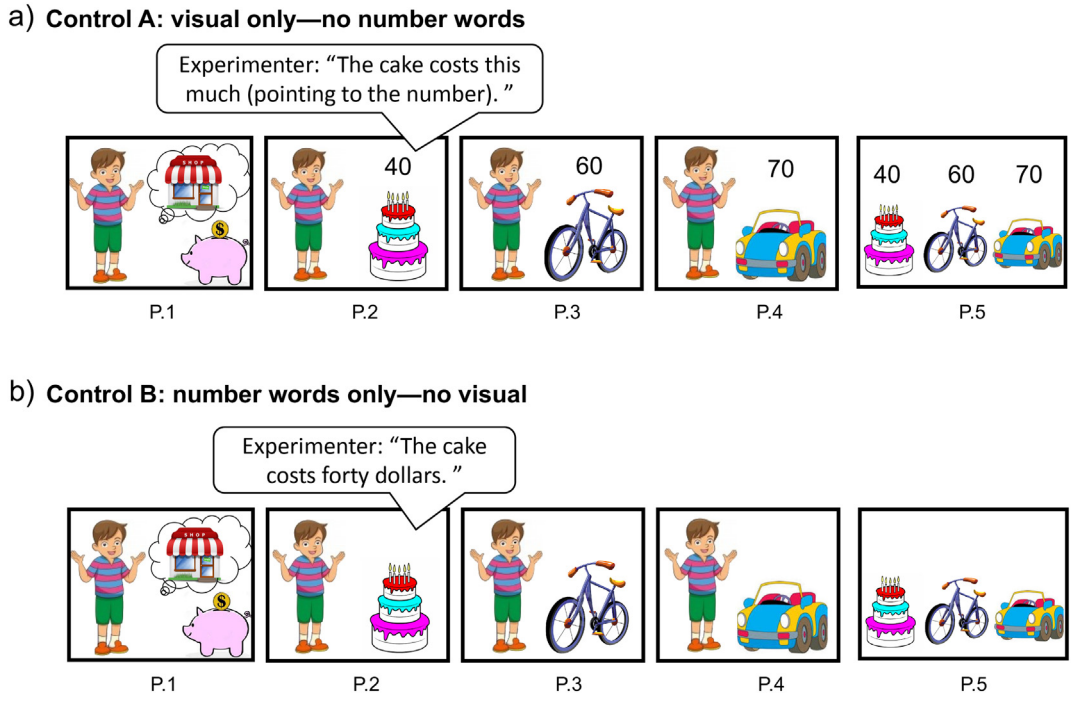

Fig. 4. Sample for the book reading activity for the two control conditions in Study 2. a) the visual only—no number words condition. Children saw pictures of objects and written numbers on the pages, but the experimenter did not provide number names during the training. b) the number words only—no visual condition. Children saw pictures of objects and heard number words from the experimenter's instruction, but never saw written numbers on the page.

Table 2
All training activities and numbers used in Experiment 2.

Sole Activity	Order	Trial	Training day 1	Training day 2	Training day 3	Total subjects
Storybook	Random	1	321	19	515	36
Storybook	Random	2	261	305	124	36
Storybook	Random	3	4	124	80	36
Storybook	Random	4	30	4	2620	36
Storybook	Random	5	421	125	4	36
Storybook	Random	6	15	205	30	36
Storybook	Random	7	570	405	14	36
Storybook	Random	8	80	1002	324	36
Storybook	Random	9	19	324	535	36
Storybook	Random	10	2	3002	24	36
Storybook	Random	11	470	6	3620	36
Storybook	Random	12	262	105	525	36
Storybook	Random	13	570	14	2	36
Storybook	Random	14	14	205	60	36
Storybook	Random	15	6	2	19	36
Storybook	Random	16	260	145	15	36
Storybook	Random	17	60	15	6	36
Storybook	Random	18	521	2002	4620	36
Make-a-number	Random	1	321	19	515	30
Make-a-number	Random	2	261	305	124	30
Make-a-number	Random	3	4	124	80	30
Make-a-number	Random	4	30	4	2620	30
Make-a-number	Random	5	421	125	4	30
Make-a-number	Random	6	15	205	30	30
Make-a-number	Random	7	570	405	14	30
Make-a-number	Random	8	80	1002	324	30
Make-a-number	Random	9	19	324	535	30
Make-a-number	Random	10	2	3002	24	30
Make-a-number	Random	11	470	6	3620	30
Make-a-number	Random	12	262	105	525	30
Make-a-number	Random	13	570	14	2	30
Make-a-number	Random	14	14	205	60	30
Make-a-number	Random	15	6	2	19	30
Make-a-number	Random	16	260	145	15	30
Make-a-number	Random	17	60	15	6	30
Make-a-number	Random	18	521	2002	4620	30

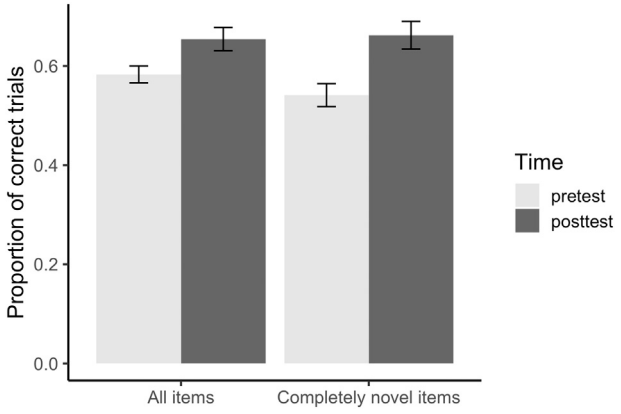

Fig. 5. Proportion of correct trials at pretest and posttest in the training condition (including both the story-book activity and the making numbers with card activity) in Study 2, for all items and for items that involved completely novel items. Error bars indicate standard errors.

novel, a Linear Mixed Effect Model (with time was entered as a fixed effect and participant was entered as a random effect) revealed a significant main effect of time, $F(1, 65) = 15.41$, $p < .001$. As shown in Fig. 5, for these totally novel test items, children's performance significantly improved from pretest ($M = 0.53$, $SE = 0.02$) to posttest ($M = 0.64$, $SE = 0.03$), $t(65) = 3.93$, $p < .001$, $d = 0.48$, providing unambiguous evidence for the generalizability of learning.

In sum, Study 2 in conjunction with Study 1 indicates that preschool children can learn and generalize the patterns that link spoken names to written multi-digit numbers. The key findings are: (1) generalization to novel numbers—numbers not experienced in training and each individually quite rare in everyday child experience (Dehaene, 1992; Gunderson & Levine, 2011; Levine et al., 2010)—is comparable or better than performance on partially experienced items and (2) this generalization did not require extensive experience with any individual items, a large sample of potential instances, or explicit instruction. We

propose that this is because the many associations that emerge from the surface structures of the multi-digit names and multi-digit numbers support systematic seemingly principled generalization.

4. Study 3

Study 3 uses a computational model to provide evidence that far generalization can result from associative learning given a data set of multiple co-predicting features that provide many overlapping and redundant pathways to the mapping between a number name and its written form. We used a form of a deep recurrent network, a general purpose associative learner, that is known to solve complex problems by exploiting multiple predictive relations (Hasson, Nastase, & Goldstein, 2019; Lecun, Bengio, & Hinton, 2015a), a fact that has led them to be criticized as un-principled, un-interpretable, and not human-like (Lake et al., 2016; Marcus, 2018). In using this general-purpose model, we make no claims that the model operates or learns in the same way as young children. Instead, the goal is to demonstrate the co-predictive properties between number names and their written forms, albeit imperfect and local predictors, are sufficient for an associative learner that does not explicitly represent any rules or principles to make far and systematic generalizations (see also, Bloom et al., 2006; Colunga & Smith, 2005; MacWhinney et al., 1989; Yoshida & Smith, 2003).

4.1. The architecture

The learning task requires linking the structure of a series of words (the number name) to the structure of an image (the written form). Specifically, on each trial of the training phase in Study 1 & 2, children were shown an image of a multi-digit number (e.g., “124”), and the experimenter provided the sequence of number names verbally (e.g., “one hundred twenty four”) while drawing children's attention to the corresponding written digits using gesture (e.g., saying “four” while pointing to “4”). Accordingly, we used an *image caption model* (Lecun et al., 2015a; Vinyals, Toshev, Bengio, & Erhan, 2015; K. Xu et al., 2015) as the algorithmic-level implementation for the proposed learning mechanism as these models are trained to generate lexical descriptions of images. Typically, these models are used to generate verbal descriptions of everyday photographs, for example, “The man in the red shirt is throwing a ball,” from an image with that content. To do this, the algorithm not only has to learn to recognize individual components of the image—e.g., objects, attributes, actions—but also their relational structure and how those relations relate to the relational structure of the lexical components of the verbal description. The computational problem is thus similar to our proposed account of how generalized knowledge of multi-digit number names and written forms might emerge. As shown in Fig. 6, we used an image caption model that is a deep neural network and has an encoder-decoder architecture with an attention mechanism (Xu et al., 2015). As described below, the encoder is used to construct a sequence of feature maps for an input image (corresponding to the images that children saw during the training), the decoder is used to generate the sequence of output words (corresponding to the sequence of number words that children heard during the training), and the attention mechanism allows the model to learn to focus on the part of the image that is most relevant to the current output word at each time step (corresponding to children's attention to individual written digits following the experimenter's gesture). All code, training, testing materials and results are available at: https://github.com/iucvl/Learning-generative-principles-of-a-symbol-system

4.1.1. Encoder

The encoder is a deep convolutional neural network (CNN) that takes an image and passes it through multiple convolution, non-linear activation, and subsampling stages. The main difference between CNNs and traditional feed-forward neural networks is that instead of fully-connected layers where each neuron is connected to all neurons in the previous layer, the network includes convolutional layers where neurons are connected to a local subset of the neurons in the previous layer. This encourages them to learn convolutional filters (e.g., 3 × 3 matrices) that extract local features (e.g., edges, textures). The subsampling stages pool features from larger spatial neighborhoods, which means that later layers of the network produce response maps that are based on evidence from larger and larger areas of the original image. In this work, we used Resnet101 (He, Zhang, Ren, & Sun, 2016), which is a particular CNN architecture that has demonstrated performance in various image classification tasks to extract features from input images. This network consists of 101 convolution and pooling layers in total. Each layer includes multiple levels of convolution followed by a non-linear activation. A pooling layer is used to reduce the size of output from the previous layers resulting in a collection of 2048 14 × 14 feature maps. These feature maps can be thought of as a mathematical representation of the abstract content of the input image (i.e., written multi-digit numbers). The network is trained using standard back propagation algorithm, in which the errors are propagated back from the output of the decoder.

4.1.2. Attention mechanism

Because a sequence of words is generated by the model to describe each visual image, the decoder needs an attention mechanism to focus on different elements in the image at each time step. We use the “soft” attention mechanism proposed by Bahdanau, Cho, and Bengio (2014): each feature map, which is reshaped into a feature vector as input to the LSTM network, is assigned a weight at each time step during decoding. The weights, updated with forward and backward propagation, are deterministic and represent the probabilities that each pixel is the place to look to generate the next word. Further details are available at: https://github.com/iucvl/Learning-generative-principles-of-a-symbol-system

4.1.3. Decoder

The decoder is a long short-term memory network (LSTM) accompanied with attention mechanism described above. This LSTM network takes a sequence, where at each time step inputs are all the feature maps (extracted by the encoder) and attention mechanism is used to decide which feature maps or parts of the feature maps are used to generate the output—in our case number names such as “three,” “hundred,” “thirty,” and “five” to describe the feature maps. LSTM is a type of recurrent neural network (RNN) frequently used for tasks that require sequence-to-sequence learning such as machine translation. RNNs contain loops in their hidden layers such that previous outputs can be used as input for the next training trial (Elman, 1990; Hochreiter & Schmidhuber, 1997). Thus, they are capable of learning the long-term dependencies among component names in a number word and the myriad correlations across different number words. LSTM networks have advantages over traditional RNNs in retaining memory of earlier time steps (Hochreiter & Schmidhuber, 1997). The network is trained with back propagation and is optimized by a loss function, which computes the cross entropy between the predicted probability (a value between 0 and 1) of the correct word and actual probability (1) for the correct word at each time step.

4.2. Training procedures

Because there is good reason to assume that most preschool children have experience with single-digit numbers, we trained the model with all single-digit numbers, including the ones that were already in the training set from Study 1 & 2. To prepare the training data for the model, we combined the training sets from Study 1 & 2. Thus, the model was given (and preliminary work showed required) more extensive training than the children (who might have had some experiences with multidigit numbers prior to the experiment). At any rate, the

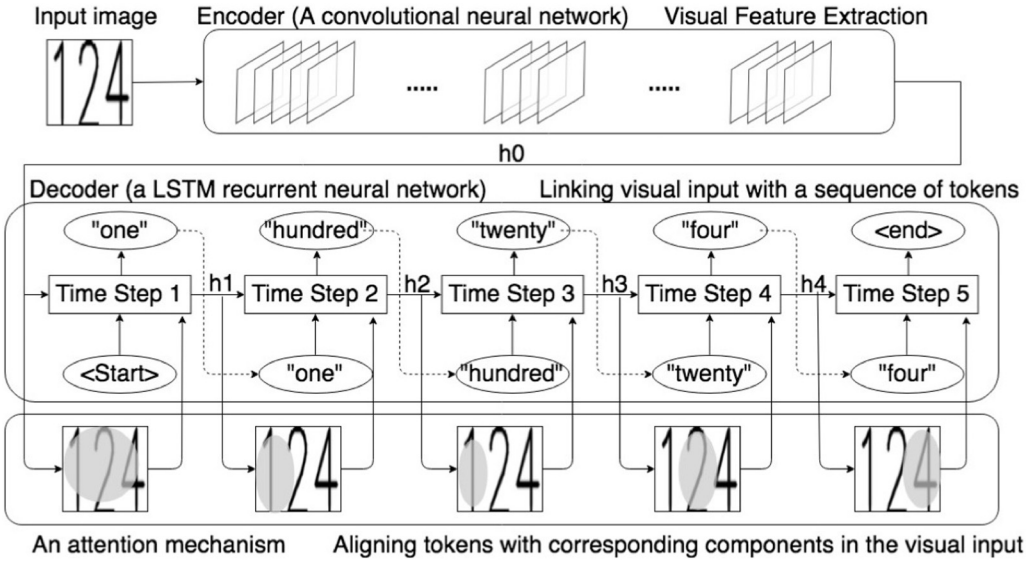

Fig. 6. Illustration of the architecture of the model, which has three basic components: 1) A convolutional neural network (CNN) as an encoder for extracting visual features from an input image, 2) A Long short-term memory (LSTM) recurrent neural network (RNN) as an decoder for linking visual input with a sequence of tokens, and 3) An attention mechanism that learns to align tokens with corresponding parts in the visual input (the shaded region represents the parts of the image that are most relevant to the token in the current time step).

final training data consisted of only 64 training trials with each trial presenting one unique number. Although larger than what was presented to children, this data set is still a quite limited sample of all the possible numbers from 1 to 9999.

The training material consisted of two streams of information—images of the written numbers and the names. For the visual information, 64 images were generated in Arial with the overall image size being constant (240 × 240 pixels). Thus, the font size of the numbers changed based on the total number of digits. This was done to prevent the model from learning based on overall size of the image (e.g., that 3-digit numbers are visually larger than 2-digit numbers). The CNN model that we used for the encoder is scale invariant; thus, changing the size of the individual digits poses no problem for this model.

Weights were initialized following a uniform distribution in the range of −0.1 to 0.1 for the decoder. The learning rate for the encoder and decoder were set to be 1e-4 and 4e-4 respectively. A total of 100 epochs were repeated for each model with the dropout rate of 0.5. Also following standard practice in the computer vision community, we pretrained this network on ImageNet (Russakovsky et al., 2015) images of everyday scenes, so that the network began learning about digits with network parameters that had some ability to represent general visual features of objects (Krizhevsky et al., 2012; Simon, Rodner, & Denzler, 2016). We ran 100 models: 50 models that were trained on pairs of input (equivalent to a post-test) and 50 models that were not trained to provide a baseline control (equivalent to a pre-test).

4.3. Testing procedures

The name of a written form was presented to the network as a sequence in time, as in spoken number names. As with all recurrent neural networks used for generating sequences of words, at each time step, the model generates a probability for each token in a library of all possible tokens. The library used in the current study included 29 tokens (see supplemental materials) that can be combined to label all numbers in the 1–9999 range. The sequence of tokens with the highest probability was taken as the number name generated by the model, and the words were combined into a final label for each input image (e.g., “one hundred twenty five” for “125”).

The children in Study 1 & 2 were tested in a two-alternative forced-choice task: given both a target number and a foil number, they needed to choose the one that matched the name. Thus, to provide comparable measures, the models were tested on all numbers in the 16 testing pairs used in Study 1 & 2. Because there are fewer constraints on testing automated models than on testing children (e.g., fatigue), and in an effort to provide more accurate estimates of the models' performance, we added 32 structurally-similar testing pairs for the model, yielding a total of 48 testing pairs (see supplemental materials for all testing items). Similar to Study 1 & 2, half of these pairs included one number (as either foil or target) number that appeared during training; for the other half of the pairs, both of the two numbers were completely novel. As described below, we provide multiple converging measures of the model's learning, from those similar to the children's forced choice task to others that probe more deeply the nature and bases of the models' performance.

4.4. Results and discussion

Measures of the accuracy of multi-component captions of images are not straightforward (Callison-Burch, Osborne, & Koehn, 2006; Papineni, Roukos, Ward, & Zhu, 2001; Vedantam, Zitnick, & Parikh, 2015). Accordingly, we used five measures that quantify performance in different but complementary ways.

4.4.1. Edit distance measure

This measure resembled the two-alternative forced-choice nature of the test, asking how similar the description provided by the model was to each of the two alternatives, with model's choice taken as the item most similar to the model's output description. We used the *Edit distance measure* that quantifies the similarity between two strings by computing the minimum number of operations required to transform one string to the other (Levenshtein, 1966). For example, to convert “eight” to “five”, we need to substitute “e” with “f,” delete “g,” “h,” and “t,” and insert “v” and “e,” resulting in a total of 6 steps. Thus, a smaller edit distance means the two strings are more similar to each other than a larger edit distance. For the purpose of the current study, for each trial (composed of a target image and a foil image), we calculate: a) the edit distance between the true label of the target image and the model-generated label based on the target image and b) the edit distance between the true label of the target image and the model-generated label based on the foil image. If either a) or b) is zero, meaning the model correctly generated the label for either or both of the target image and foil image, we scored the model as correct. If neither a) nor b) is zero, but a) $<$ b), we scored the model as correct. If neither a) or b) is zero, and a) $>$ b), we scored the model as incorrect.

Fig. 7 (a) shows the edit distance measure for the untrained models (n = 50) and the trained models (n = 50) after 100 iterations. A linear mixed effect model (LMM) was conducted in which time was entered as

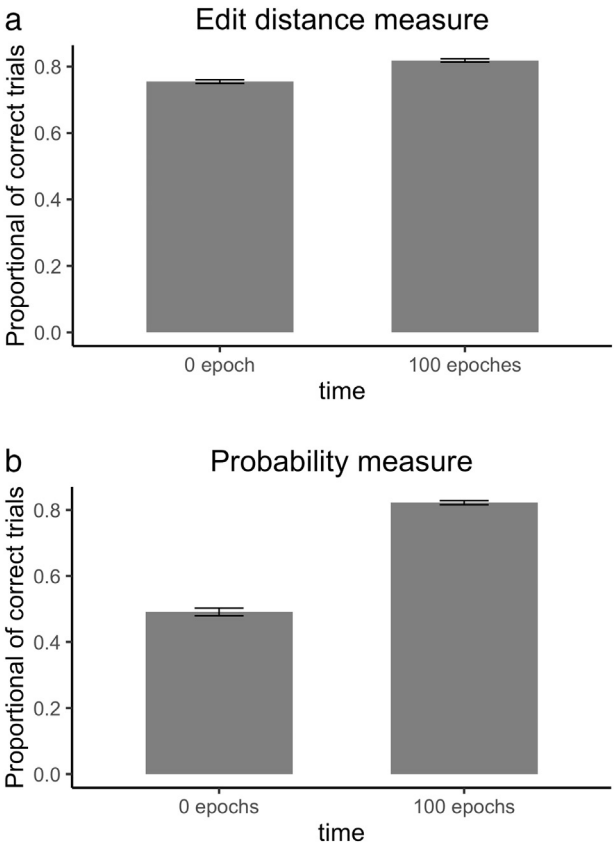

Fig. 7. Model performance based on (a) the edit distance measure (b) the probability measure after 0 epoch and 100 epochs. Error bars indicate standard errors.

a fixed effect and model identity was entered as a random effect. Results showed a significant main effect of training, F (1, 49) = 170.79, p < .0001. The trained models performed significantly better (M = 0.82, SE = 0.005) than the untrained models (M = 0.76, SE = 0.005) in their ability to provide a label that was more similar to the correct number name than to that of the foil, t (49) = 13, p < .0001, d = 1.84. A separate linear mixed effect model (LMM) was conducted—on only test items containing numbers that never occurred in the training—with time as a fixed effect and model (trained versus untrained) as a random effect yielded a significant main effect of training, F (1,49) = 38.75, p < .0001. The trained models performed significantly better (M = 0.77, SE = 0.01) than the untrained models (M = 0.72, SE = 0.01) in their ability to choose a written untrained number between two choices given a number name, t (49) = 6.20, p < .0001, d = 0.88.

4.4.2. Probability measure

The model outputs a distribution of probabilities for all tokens (the components of number names) in the library at each time step. We computed the average probability for the correct tokens as follows. Suppose the current trial includes two numbers “78” (the target) and “260” (the foil), with the desired label being “seventy eight.” For the target image, at time step 1, we took the model-generated probability of the token “seventy” (P1) (regardless of whether that token had the highest probability). At time step 2, we took the probability for the token “eight” (P2) (regardless of whether that token had the highest probability). The overall probability for the target image was then computed by averaging P1 and P2. This number can be interpreted as how probable the model thinks that the target image should be named by the desired label. Similarly, for the foil image, at time step 1, we took the model-generated probability of the token “seventy” (P3) and at time step 2, we took the probability for the token “eight” (P4). The overall probability for the foil image was then computed by averaging P3 and P4, and can be interpreted as how probable the model thinks that the foil image should be named by the desired label. If the overall probability of the target image was higher than that of the foil image, then the current trial was scored as correct.

A linear mixed effect model (LMM) with time entered as a fixed effect and model identity entered as a random effect yielded a significant main effect of training, F (1, 49) = 771.05, p < .0001. As shown in Fig. 7(b), the trained models performed significantly better (M = 0.82, SE = 0.01) than the untrained models (M = 0.49, SE = 0.01) in their ability to choose a written number between two choices given a number name, t (49) = 28, p < .0001, d = 3.92. Performance on test items that involved completely novel target and foil numbers was examined in a Linear Mixed Effect model (LMM) with time was entered as a fixed effect and model identity as a random effect and yielded a significant main effect of training, F (1, 49) = 265.53, p < .0001. The trained models performed significantly better (M = 0.83, SE = 0.01) than the untrained models (M = 0.49, SE = 0.02) as measured by a better matching output description for the target than the foil, t (49) = 16, p < .0001, d = 2.3.

4.4.3. Correlation measure

The first two measures assess the relative similarity of the output number name for the target versus the foil. But one can also ask how well the generated name captures correct components of the target, even if not totally correct. For example, if the model's output for the numbers “256” and “147” are “two hundred fifty” and “one hundred seven,” the model would seem to have partial knowledge of how names map to written forms. At the very least, the outputs preserve the ordinal relation between the numbers (256 > 147, “two hundred fifty” > “one hundred seven”). One way to capture this is to compute the correlation between the numerical values of the true labels and the numerical values of the generated labels (Yuan et al., 2019). We did this for all input images, target and foil (n = 4800). The generated value and the true value were highly correlated, Spearman correlation r = 0.83, p < .0001.

4.4.4. Attention measure

Image caption models have to learn where to look in a scene. If the models have learned how the temporal order of elements in the name corresponds to the spatial elements of the written form, they should show systematic biases in how the temporal sequences of tokens are related to the attended spatial locations in the input image. That is, the model should “inspect” the image from left-to-right while producing the number words. Accordingly, we calculated the probability that the model was “attending” to the left versus right side of the image, when the first word versus the last word of the label was outputted by the model. As expected, the models were more likely to attend to the left side of the image when the first word was “spoken” (56% vs 44%), but more likely to attend to the right side of the image when the last word was “spoken” (31% vs 69%), excluding single digit numbers and numbers composed of only one word (17% of total data).

4.4.5. Error patterns

To provide further evidence that it is the overlapping surface predictors that are the basis of the networks and the children's far generalization, we examined the kinds of item types on which models and children (in Study 1 and Study 2) were most likely to make errors. If the model and the children were generalizing on the basis of the same kinds of partial local predictive relations, they should show a pattern of errors predictable by overlapping predictors that match versus distinguish the target from the foil. Past research (Yuan et al., 2019) on children's errors classified the relation between target and foil into four mutually exclusive categories: single digit numbers (S, e.g., 2 vs 8) which have no overlapping predictors between the name and the form, multi-digit

numbers with different numbers of places (M-DP, e.g., 25 vs 405) which can be discriminated by predictors such as “hundred” and “-ty” as well as individual components, multi-digit numbers with the same number of places but no transposition (M-SP-no-T, e.g., 608 vs 658) which can be discriminated by at least one spoken name to digit (e.g., “fifty” predicting 5), and multi-digit numbers with the same number of places and transpositions (M-SP-T, e.g., 306 vs 360). Success on this last type of items requires the simultaneous application of more predictive elements. For example, to solve items that are multi-digit numbers with the same number of places and transpositions (e.g., 306 versus 360)—M-SP-T—the model or the child has to know the precise mapping between place value terms and the individual digits in a number but that the symbol “0” does not get named, that “hundred” signals the “3” in “306”, and that the temporal sequence of number words corresponds to the spatial location from left-to-right in the written form. Items in the other categories may be solved with just one or several predictive components. For example, to figure out which number is “twenty five” in the pair of “25” and “405”, the child or the network may rely on any of these associations—that “twenty” refers to numbers with a “2”, that three-digit numbers must have the word “hundred” in its name, that “4” corresponds to “four” in the name, and so on. If this analysis is correct, then children and the networks should perform most poorly in the M-SP-T category and better on M-DP and M-SP-no-T categories. Performance on mapping number names to single digits (which is required along with other associations on the other items) should yield the best performance. We used both the edit distance and the probability measure to assess the networks' performances. As shown in Fig. 8, the neural networks and children in Study 1 and 2 showed the same *ordinal pattern* of errors, consistent with their use of the same kinds of information. Clearly, there are also differences suggesting that the children and the model may weight different predictive factors differently based on prior experience or that mechanisms at the response stage influence children's behaviors.

Overall, the above five measures provide converging evidence that (1) learned associations between a limited sampling of multi-digit number names and written forms are sufficient to yield knowledge about how number names *in general* map to written multi-digit numbers, and (2) that learning about multiple and local predictive relations between surface properties of names and numbers leads to far generalization. These findings from the model provide additional support for our main conclusion: Data sets with several local predictors and thus many paths to generalization lead to rapid learning and systematic generalization from just a few examples.

5. General discussion

Trained with just 36 unique numbers and their names and with just one exposure for most of the numbers and names, preschool children mapped multi-digit names to their written forms for instances not experienced in the training, instances that were also individually unlikely to have been encountered in everyday experiences. Studies 1 and 2 used two different randomly selected training sets, suggesting that the particular training items do not matter, and that many different samples of numbers across the range 1 to 9999 would be effective. Studies 1 and 2 used two different training contexts—in combination and alone—yielding the same outcomes and suggesting that the particular training context in which the names and written forms co-occur is also not critical. Presented with a slightly larger training set (64 unique items) and no repetitions, the model in Study 3 also performed well, generating number names given images of untrained multi-digit numbers. Previous research indicates that preschool children have minimal understanding of the actual meaning of places (Fuson, 1990; Mix, Smith, & Crespo, 2019; Ross, 1995), and many school-aged children as late as 5th grade (Ross, 1986; Ross & Sunflower, 1995) still struggle to understand the multiplicative hierarchy that underlies base-10 notation. Thus, it is highly unlikely that the minimal training in Studies 1 and 2 taught the children the meaning of the places, for example, that the “4” in “346” represents 4 sets of 10. Rather, children's generalizations to novel instances likely reflect the acquisition and exploitation of a myriad of predictive relations between names and written forms: for example, that “two hundred fifty-six” as well as “forty-two” predict a “2” somewhere in the written form, that “two hundred fifty-two” predicts two “2's” with one in the left-most position, that “twenty” and

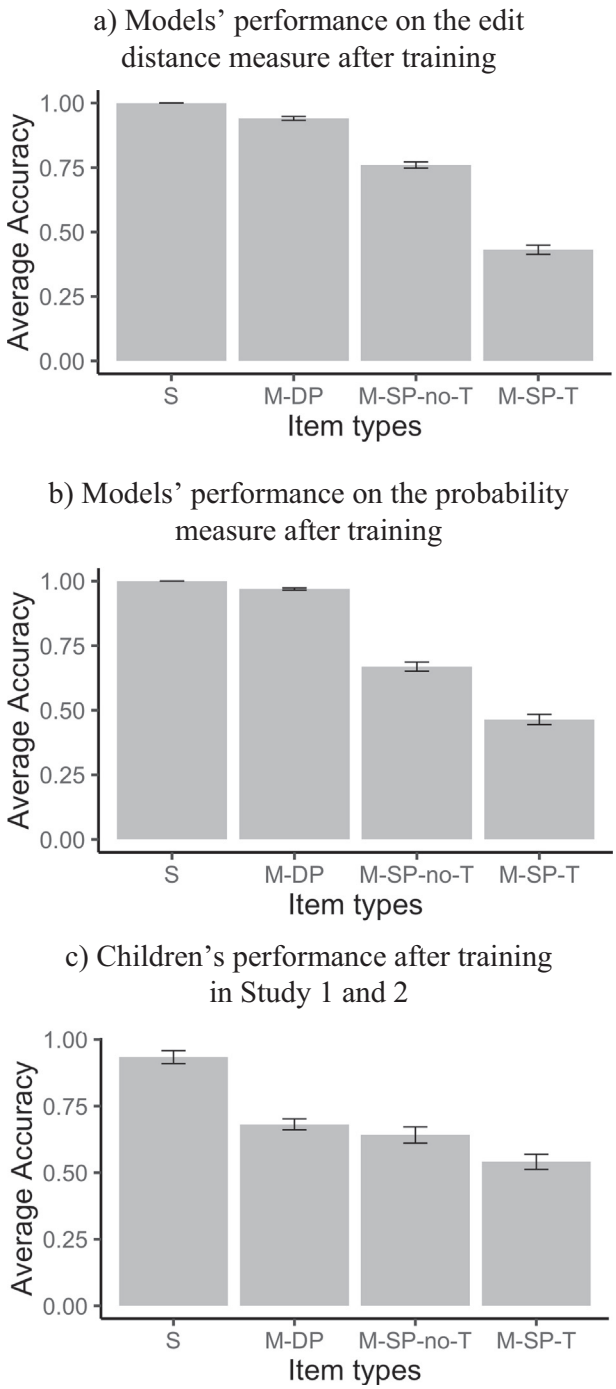

Fig. 8. Error patterns across different types of testing items—S = single digit numbers, M-DP = multi-digit numbers with different numbers of places, M-SP-no-T = multi-digit numbers with the same number of places but no transpositions, M-SP-T = multi-digit numbers with the same number of places and transpositions—for models' performance on the edit distance measure after training (a), models' performance on the probability measure after training (b), and children's performance after training in Study 1 and 2 (c). Error bars indicate standard errors.

"two" both predict a "2" in the written form, and so forth. There is good reason to believe that the network models in Study 3 succeeded on a similar basis (LeCun, Bengio, & Hinton, 2015b; Rumelhart, Hinton, & Williams, 1985; Schmidhuber, 2015). Thus, the entire pattern of results provides evidence for systematic and broad generalization from a relatively few training instances that is not dependent on the explicit learning or representation of the underlying generative principles.

In her earlier work on this idea, Billman (Billman, 1989, 1993; Billman & Knutson, 1996) used the term *systematicity* to refer to multiple inter-predictive features that offer *multiple* pathways to generalization. This kind of systematicity in a data set appears to be readily exploitable by human learners and may characterize a variety of knowledge domains shared by many individuals—not just the surface properties of place value notation but also many aspects of language (Bloom et al., 2006; Christiansen & Monaghan, 2010; MacWhinney et al., 1989), as well as superordinate-level and basic-level object categories (McMurray et al., 2012; Rogers & McClelland, 2004; Rosch, 1978; Samuelson, 2002). In these cases, individual learners can have quite idiosyncratic experiences—specific to their personal history—and yet generalize and generate patterns consistent with other learners. This robustness, as well as the ability to generalize from relatively few experiences, may emerge because such Billman-style systematicity builds on many overlapping inter-predictive features with most encountered instances (with some exceptions such as "eleven") presenting at least some of these many predictive features. Knowledge domains learned and used by many individuals may evolve to increase local overlapping predictive patterns precisely because human learners are sensitive to and readily exploit them (Christiansen & Kirby, 2003; Kirby, Griffiths, & Smith, 2014; Monaghan, Shillcock, Christiansen, & Kirby, 2014). Thus, in many domains of human cognition, a hodgepodge of multiple inter-predictive features may be sufficient to account for human generalization (Colunga & Smith, 2005; Hasson et al., 2019; MacWhinney et al., 1989; Seidenberg & McClelland, 1989).

The children (and the model) in the present studies were purposely given minimal training to make the point that this kind of systematicity is easily found and generalized by learners. Although the findings show clear evidence for this conclusion, performance was well below mastery. There could be more dramatic generalizations given further experiences with number names and their written forms. If one examines the patterns of overlapping associations in Fig. 1, one can see that two spoken elements form a hub, "-ty" and "hundred" which mark the places and form categories of the digits that fall in them. If the example network had included 4-digit numbers, "thousand" would also be part of that hub. Given this structure, more experiences and a well-entrenched learning of the inter-predictive patterns for 1 to 9999, children (and deep learning networks) might well show one-shot learning (extrapolation) beyond that range: exposure to the name and written form for just one novel number, such as 21,578, might be sufficient for the learner to *generate* number names for any number from 1 to 99,999. The learning and explicit representation of well-formed rules and principles often seems to be the pinnacle of human learning (Lake et al., 2015; Rule et al., 2015; Tenenbaum et al., 2011), but a great deal of human intelligence could rest on exploiting a plethora of local, inter-predictive, and imperfect surface features. This in-principle possibility is well-demonstrated in contemporary machine learning.

However, to succeed in the basics of arithmetic, young children must go beyond the predictive patterns in the surface forms to understand the meaning of places and the multiplicative hierarchy that provides that meaning. How might learning about the inter-predictive surface patterns be related to learning about the generative principles? Many related variants of this question populate the literature in cognitive science: implicit versus explicit learning (Reber, 1989), associative versus propositional representations (Chomsky, 1980; Fodor & Pylyshyn, 1988), intuitive processes versus conscious rule interpreters (Smolensky, 1988), intuitive and rational processing (Hinton, 1990), associative versus rule-based reasoning (Sloman, 1996), connectionism versus probabilistic reasoning (Griffiths et al., 2010; McClelland et al., 2010), statistical learning versus hypothesis testing (Medina, Snedeker, Trueswell, & Gleitman, 2011; Smith & Yu, 2008). We believe that children's learning about base-10 notation provides a rich and well-defined context within which to make progress on these inter-related issues. With this larger goal in mind, we offer two hypotheses about how learning the multiple predictive patterns in the surface structure of number names and written forms may be related to learning the principles underlying the multiplicative hierarchy of places.

One possibility is that these are fundamentally distinct forms of knowledge. Nonetheless, the early learning of predictive patterns relating number names and written multi-digit forms may support learning the principles of base-10 notation by guiding in-the-moment attentional processes during formal instruction (Yuan et al., 2019). A large literature shows that known words automatically direct attention to referents in crowded visual fields (Huettig & McQueen, 2007; Lupyan & Ward, 2013; Spivey, Tyler, Eberhard, & Tanenhaus, 2001). Formal in-school instruction about the multiplicative hierarchy often occurs in highly cluttered contexts of number lines and number boards with many written multi-digit numbers in view, heard number talk, and grounding activities such as the bundling and unbundling of physical sets of 10. To learn, children must look to relevant visual information at the right moment. The modeling results of Study 3 show that the surface regularities in number names and corresponding written forms are sufficient for the internal *components* of spoken number names to direct attention to the spatial regions *within* a multi-digit number. This facility in looking behavior—acquired through learned associations between the number names and written forms—may enable children to more accurately attend to the components of a string of digits and enable them to connect the relevant components to each other and to grounding activities about sets of 10. In so doing, early learning of the partial inter-predictive mappings between names and written forms may prevent the formation of wrong ideas that characterize some children's knowledge of the place value system even as late as 5th grade (Gervasoni et al., 2011; Ross & Sunflower, 1995). This possibility has direct and actionable implications for understanding why some children falter in learning about the place value system while other children—in the same classrooms—readily succeed (Yuan et al., 2019).

The second possibility is that that the key grounding for learning about base-10 notation lies not in the world and concrete examples of bundled sets of 10 sticks, but in the latent structure of many predictive correlations within the symbol system itself, the latent knowledge apparent in the hub at the center of network of surface-level associations (Fig. 1). Advanced associative models can find higher-order correlations that represent abstract categories such as nouns and verbs, or the distinction between mass nouns and count nouns (Colunga & Smith, 2005; Landauer & Dumais, 1997; Rogers & McClelland, 2004). Image captioning algorithms, like that used in Study 3, trained on visual scenes and sentences describing those scenes have sufficient latent knowledge to generate grammatically-correct sentences without any training on syntactic categories (Datta et al., 2019; Xu et al., 2015). Thus, children's learning of many overlapping inter-predictive features between the surface properties of names and the written symbols may form the internal knowledge of places that is made explicit with formal training, just as training in grammar brings forward explicit knowledge about nouns phrases and verb phrases. Ultimately, the understanding of place value requires an understanding of the relational structure among the places; forming latent categories of places may be essential for such an explicit understanding of the multiplicative hierarchy. This idea that meaning of places originates in the latent structure inherent in the surfaces features of the symbol system may explain why fully generative principles can be approached but perhaps not fully realized for human learners, as evident in the limits on many adults' understanding of the base-10 notation when confronted with very large numbers (e.g., millions and billions, Landy, Charlesworth, & Ottmar, 2016; Landy, Silbert, & Goldin, 2013).

Each of these two possibilities—being facile with the symbols benefits explicit learning about place value and early statistical learning teaches the underlying relational structure—are not mutually exclusive and both require more extensive empirical study. The contribution of the present work to research on education is three-fold: First, it highlights a potential educational role for statistical learning from mere experience—without tasks, explanations or feedback. Second, it offers an origin for and potential solution to the problem of why some children succeed and others falter from the same formal instruction about place value. Some children may have discovered the statistical regularities behind these mappings long before school; and finally, the results suggest a new agenda for research on early mathematics education, one that focuses less on grounding and explanation of abstract concepts and more on how learners form latent knowledge about the symbol systems that affects future learning.

In conclusion, learning depends on the internal mechanisms, the structure of the learning domain, and the prior learning of the learner. The study of children's learning about the place value system offers a complex and tractable domain within which to make progress on how all these components fit together. Critical to this progress is the study of the data structures that characterize real world learning problems as they naturally occur. Growing evidence suggests that the statistical structure of everyday experience often differs fundamentally from the kinds of data structures used in laboratory experiments of human learning and those used to train machine learning models (Bambach, Crandall, Smith, & Yu, 2018; Dupoux, 2018; Frankenhuis, Nettle, & Dall, 2019; Smith & Slone, 2017). Current deep learning models are commonly criticized as data hungry and as being able to learn only local similarities—not general principles—despite all that data (Feinman & Lake, 2018; Lake et al., 2016; Marcus, 2018). The present findings suggest a more complete and unified understanding of all forms of learning and their relation to each other might best begin by studying natural real-world data sets for real world learning problems.

Note. The input to the LSTM network and on which the attention mechanism operates was the series of localized feature maps extracted by the CNN. But, for the ease of interpretation, the raw input image was shown at the bottom to demonstrate the final learning outcome of the attention mechanism where the network over time learned to prioritize the most relevant part of the image for predicting the current component word.

CRediT authorship contribution statement

Lei Yuan: Conceptualization, Methodology, Investigation, Formal analysis, Validation, Visualization, Writing - original draft. **Violet Xiang:** Software, Data curation, Writing - original draft. **David Crandall:** Supervision, Resources, Writing - review & editing. **Linda Smith:** Conceptualization, Methodology, Writing - review & editing.

Acknowledgements

This research was supported by NICHD F32 HD090827-02 grant to Lei Yuan, NSF DRL 1621-93 grant to Linda Smith and Kelly Mix, and the Indiana University Emerging Areas of Research (EAR) grant. We thank Emily Johns and Haley Meekhof for help with data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104243.

References

Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(1–2), 1–7. https://doi.org/10.1002/wcs.1373.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate.* arXiv preprintarXiv:1409.0473.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in neural information processing systems. Vol. 2018-Decem. Advances in neural information processing systems* (pp. 1201–1210).

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Bolker, M. B. (2015). Package 'lme4'. *Convergence, 12*(), 2.

Berko, J. (1958). The child's learning of English morphology. *Word,* 150–177 14 SRC-.

Billman (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language and Cognitive Processes, 4*(2), 127–155.

Billman, D. (1993). Influences from multiple functions. In D. L. Medin (Ed.). *Psychology of learning and motivation: Advances in research and theory* (pp. 283). Academic Press.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*(2), 458–475. https://doi.org/10.1037/0278-7393.22.2.458.

Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30months. *Cognition, 126*(1), 39–53. https://doi.org/10.1016/j.cognition.2012.08.008.

Bloom, L., Lightbown, P., Hood, L., Bowerman, M., Maratsos, M., & Maratsos, M. P. (2006). Structure and variation in child language. *Monographs of the Society for Research in Child Development, 40*(2), 1. https://doi.org/10.2307/1165986.

Botvinick, M., & Plaut, D. C. (2004). Doing without Schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review, 111*(2), 395–429. https://doi.org/10.1037/0033-295X.111.2.395.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Byrge, L., Smith, L. B., & Mix, K. S. (2014). Beginnings of place value: How preschoolers write three-digit numbers. *Child Development, 85*(2), 437–443. https://doi.org/10.1111/cdev.12162.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. *EACL 2006 - 11th conference of the European Chapter of the Association for Computational Linguistics, proceedings of the conference* (pp. 249–256). .

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *The Stanford Child Language Development. Vol. 15. The Stanford Child Language Development* (pp. 17–29).

Casler, K., & Kelemen, D. (2005). Young children's rapid learning about artifacts. *Developmental Science, 8*(6), 472–480. https://doi.org/10.1111/j.1467-7687.2005.00438.x.

Chi, M. T. H., Kristensen, A. K., & Roscoe, R. D. (2012). Misunderstanding emergent causal mechanism in natural selection. *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 145–173). . https://doi.org/10.1093/acprof:oso/9780199730421.003.0007.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences, 3*(1), 1–15. https://doi.org/10.1017/S0140525X00001515.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences, 7*(7), 300–307. https://doi.org/10.1016/S1364-6613(03)00136-0.

Christiansen, M. H., & Monaghan, P. (2010). Discovering verbs through multiple-cue integration. *Action meets word: How children learn verbs*https://doi.org/10.1093/acprof:oso/9780195170009.003.0004.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review, 112*(2), 347–382. https://doi.org/10.1037/0033-295X.112.2.347.

Cooper, L. L., & Tomayko, M. C. (2011). Understanding place value. *Teaching Children Mathematics, 17*(9), 558–567.

Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., & Divakaran, A. (2019). Align2Ground: Weakly supervised phrase grounding guided by image-caption alignment. *Computer vision and pattern recognition*.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition.*. https://doi.org/10.1016/0010-0277(92)90049-N.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition, 43*(1), 1–29. https://doi.org/10.1016/0010-0277(92)90030-L.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition, 173*, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1.

Fe-Fei, Fergus, & Perona (2003). A Bayesian approach to unsupervised one-shot learning of object categories. *Proceedings ninth IEEE international conference on computer vision. vol. 2. Proceedings ninth IEEE international conference on computer vision* (pp. 1134–1141). IEEE. https://doi.org/10.1109/ICCV.2003.1238476.

Feinman, R., & Lake, B. M. (2018). *Learning inductive biases with simple neural networks*.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.

Frankenhuis, W. E., Nettle, D., & Dall, S. R. X. (2019). A case for environmental statistics of early-life effects. *Philosophical Transactions of the Royal Society B: Biological Sciences, 374*(1770), 20180110. https://doi.org/10.1098/rstb.2018.0110.

Fuson, K. C. (1986). Roles of representation and verbalization in the teaching of multi-digit addition and subtraction. *European Journal of Psychology of Education, 1*(2), 35–56. https://doi.org/10.1007/BF03172568.

Fuson, K. C. (1988). *Children's counting and concepts of number.* New York, NY, US:

Springer-Verlag Publishing.
Fuson, K. C. (1990). Conceptual structures for multiunit numbers: Implications for learning and reaching multidigit addition, subtraction, and place value. *Cognition and Instruction, 7*(4), 343–403. https://doi.org/10.1207/s1532690xci0704_4.
Fuson, K. C., & Briars, D. J. (1990). Using a base-ten blocks learning/teaching approach for first- and second-grade place- value and multidigit addition and subtraction. *Journal of Research in Mathematics Education, 21*(3), 180–206.
Fuson, K. C., & Kwon, Y. (1991). Chinese-based regular and European irregular systems of number words: The disadvantages for English-speaking children. *Language in mathematical education: Research and practice* (pp. 211–226). .
Fuson, K. C., & Kwon, Y. (1992). Korean children's understanding of multidigit addition and subtraction. *Child Development, 63*(2), 491–506. https://doi.org/10.2307/1131494.
Geary, D. C., Bow-Thomas, C. C., Liu, F., & Siegler, R. S. (1996). Development of arithmetical competencies in Chinese and American children: Influence of age, language, and schooling. *Child Development, 67*(5), 2022–2044. https://doi.org/10.1111/j.1467-8624.1996.tb01841.x.
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., Levine, S. C., Ping, R., Isaia, A., Dhillon, S., Bradley, C., & Honke, G. (2016). Rapid learning in a children's museum via analogical comparison. *Cognitive Science, 40*(1), 224–240. https://doi.org/10.1111/cogs.12248.
Gervasoni, A., Parish, L., Hadden, T., Turkenburg, K., Bevan, K., Livesey, C., & Croswell, M. (2011). Insights about children's understanding of 2-digit and 3-digit numbers. *Mathematics traditions and new practices proceedings from the Australian Associat*.
Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 75–86. https://doi.org/10.1002/wcs.1330.
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004.
Grossberg, S., & Repin, D. V. (2003). A neural model of how the brain represents and compares multi-digit numbers: Spatial and categorical processes. *Neural Networks, 16*(8), 1107–1140. https://doi.org/10.1016/S0893-6080(03)00193-X.
Gunderson, E. A., & Levine, S. C. (2011). Some types of parent number talk count more than others: Relations between parents' input and children's cardinal-number knowledge. *Developmental Science, 14*(5), 1021–1032. https://doi.org/10.1111/j.1467-7687.2011.01050.x.
Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning difficulties. *Journal of Educational Psychology, 93*(3), 615–626. https://doi.org/10.1037/0022-0663.93.3.615.
Hasson, U., Nastase, S. A., & Goldstein, A. (2019). Robust-fit to nature: An evolutionary perspective on biological (and artificial) neural networks. *BioRxiv, 764258*. https://doi.org/10.1101/764258.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition. Vol. 2016-Decem. Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition* (pp. 770–778). . https://doi.org/10.1109/CVPR.2016.90.
Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence, 46*(1–2), 47–75. https://doi.org/10.1016/0004-3702(90)90004-J.
Ho, C. S.-H., & Fuson, K. C. (1998). Children's knowledge of teen quantities as tens and ones: Comparisons of Chinese, British, and American kindergartners. *Journal of Educational Psychology, 90*(3), 536–544. https://doi.org/10.1037/0022-0663.90.3.536.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*(4), 460–482. https://doi.org/10.1016/J.JML.2007.02.001.
Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development, 9*(1), 45–75. https://doi.org/10.1016/0885-2014(94)90019-1.
Kamii, C. (1986). Place value: An explanation of its difficulty and educational implications for the primary grades. *Journal of Research in Childhood Education, 1*(2), 75–86. https://doi.org/10.1080/02568548609594909.
Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). *Theory acquisition and the language of thought*.
Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 103*, 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x.
Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology, 28*, 108–114. https://doi.org/10.1016/J.CONB.2014.07.014.
Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*(6), 2797–2822. https://doi.org/10.1111/j.1467-8624.1996.tb01889.x.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*.
Lake, B. M., Linzen, T., & Baroni, M. (2019). *Human few-shot learning of compositional instructions*.
Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences, 2012*, 1–101. https://doi.org/10.1017/S0140525X16001837.
Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*(3), 299–321. https://doi.org/10.1016/0885-2014(88)90014-7.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211.
Landy, D., Charlesworth, A., & Ottmar, E. (2016). Categories of large numbers in line estimation. *Cognitive Science,* 1–28. https://doi.org/10.1111/cogs.12342.
Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science, 37*(5), 775–799. https://doi.org/10.1111/cogs.12028.
Lecun, Y., Bengio, Y., & Hinton, G. (2015a). Deep learning. *Nature*Nature Publishing Grouphttps://doi.org/10.1038/nature14539.
LeCun, Y., Bengio, Y., & Hinton, G. (2015b). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.
Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.
Levine, S. C., Suriyakham, L. W., Rowe, M. L., Huttenlocher, J., & Gunderson, E. A. (2010). What counts in the development of young children's number knowledge? *Developmental Psychology, 46*(5), 1309–1319. https://doi.org/10.1037/a0019671.
Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y.
Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences of the United States of America, 110*(35), 14196–14201. https://doi.org/10.1073/pnas.1303312110.
MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28*(3), 255–277. https://doi.org/10.1016/0749-596X(89)90033-8.
Mann, A., Moeller, K., Pixner, S., Kaufmann, L., & Nuerk, H. C. (2012). On the development of Arabic three-digit number processing in primary school children. *Journal of Experimental Child Psychology, 113*(4), 594–601. https://doi.org/10.1016/j.jecp.2012.08.002.
Marcus, G. (2018). *Deep learning: A critical appraisal.* arXiv preprintarXiv:1801.00631.
McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14*(8), 348–356. https://doi.org/10.1016/j.tics.2010.06.002.
McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review, 119*(4), 831–877. https://doi.org/10.1037/a0029872.
Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences of the United States of America, 108*(22), 9014–9019. https://doi.org/10.1073/pnas.1105040108.
Mervis, C. B., & Johnson, K. E. (1991). Acquisition of the plural morpheme: A case study. *Developmental Psychology, 27*(2), 222–235. https://doi.org/10.1037/0012-1649.27.2.222.
Miura, I. T., & Okamoto, Y. (1989). Comparisons of U.S. and Japanese first graders' cognitive representation of number and understanding of place value. *Journal of Educational Psychology, 81*(1), 109–114. https://doi.org/10.1037/0022-0663.81.1.109.
Mix, K. S., Prather, R. W., Smith, L., & Stockton, J. D. (2014). Young children's interpretation of multidigit number names: From emerging competence to mastery. *Child Development, 85*(3), 1306–1319. https://doi.org/10.1111/cdev.12197.
Mix, K. S., Smith, L. B., & Crespo, S. (2019). Leveraging relational learning mechanisms to improve understanding of place value. *Constructing number - Merging perspectives from psychology and mathematics education* (pp. 87–121). Springer Publishing. https://doi.org/10.1007/978-3-030-00491-0_5.
Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130299. https://doi.org/10.1098/rstb.2013.0299.
Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU. *Proceedings of the 40th annual meeting on association for computational linguistics - ACL '02* (pp. 311). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.
Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219.
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* MIT Press.
Rosch, E. (1978). Principals of categorization. *Concepts: Core readings* (pp. 251–270). .
Ross (1986). The development of children's place-value numeration concepts in grades two through five. *Annual meeting of the American Educational Research Association*https://doi.org/10.1017/CBO9781107415324.004.
Ross (1995). *Children's acquisition of place-value numeration concepts: The roles of cognitive development and instruction. Focus on learning problems in mathematics. 12*, 1 SRC-Google Scholar FG-01–17.
Ross, & Sunflower (1995). *Place-value: Problem-solving and written assessment using digit-correspondence tasks. In the 17th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*OH: Columbus.

Rule, J., Dechter, E., & Tenenbaum, J. B. (2015). Representing and learning a large system of number concepts with latent predicate networks. *CogSci,* 2051–2056.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y.
Samuelson, L. K. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15–20-month-olds. *Developmental Psychology, 38*(6), 1016–1037. https://doi.org/10.1037/0012-1649.38.6.1016.
Saxton, M., & Towse, J. (1998). Linguistic relativity: The case of place value in multi-digit numbers. *Journal of Experimental Child Psychology, 69*(1), 66–79. https://doi.org/10.1006/jecp.1998.2437.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117. https://doi.org/10.1016/J.NEUNET.2014.09.003.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523.
Siegler, R. S., & Lortie-Forgues, H. (2017). Hard lessons: Why rational number arithmetic is so difficult for so many people. *Current Directions in Psychological Science, 26*(4), 346–351. https://doi.org/10.1177/0963721417700129.
Simon, M., Rodner, E., & Denzler, J. (2016). *ImageNet pre-trained models with batch normalization*.
Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2015). *afex: Analysis of factorial experiments. R package version 0.13–145*.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3.
Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition, 60*(2), 143–171. https://doi.org/10.1016/0010-0277(96)00709-3.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*(1), 13–19. https://doi.org/10.1111/1467-9280.00403.
Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology, 42*(6), 1339–1343. https://doi.org/10.1037/0012-1649.42.6.1339.
Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.02124.
Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010.
Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*(1), 1–23. https://doi.org/10.1017/S0140525X00052432.
Son, J. Y., Smith, L., & Goldstone, R. L. (2012). Connecting instances to promote children's relational reasoning. *Journal of Experimental Child Psychology, 108*(2), 323–343. https://doi.org/10.1016/j.jecp.2010.08.011.Connecting.
Spelke, E. S. (2016). Core knowledge and conceptual change: A perspective on social cognition. In D. Barner, & B. A. Scott (Eds.). *Core knowledge and conceptual change* (pp. 279–300). NY: New York: Oxford University Press.
Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science, 12*(4), 282–286. https://doi.org/10.1111/1467-9280.00352.
Team, R. C (2017). *R: A language and environment for statistical com-puting.* Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788.
Treiman, R. (1993). *Beginning to spell: A study of first-grade children.* Oxford University Press.
Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4566–4575). IEEE. https://doi.org/10.1109/CVPR.2015.7299087.
Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3156–3164). IEEE. https://doi.org/10.1109/CVPR.2015.7298935.
Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). *Galileo: Perceiving physical object properties by integrating a physics engine with deep learning*.
Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245.
Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning* (pp. 2048–2057). .
Yoshida, H., & Smith, L. B. (2003). Correlation, concepts and cross-linguistic differences. *Developmental Science, 6*(1), 30–34. https://doi.org/10.1111/1467-7687.00249.
Yuan, L., Prather, R. W., Mix, K. S., & Smith, L. B. (2019). Preschoolers and multi-digit numbers: A path to mathematics through the symbols themselves. *Cognition, 189*, 89–104. https://doi.org/10.1016/j.cognition.2019.03.013.
Yuan, L., Uttal, D., & Gentner, D. (2017). Analogical processes in children's understanding of spatial representations. *Developmental Psychology, 53*(6), 1098–1114. https://doi.org/10.1037/dev0000302.