How children learn to value numbers: Information structure and the acquisition of numerical understanding

Michael Ramscar, Melody Dye, Hanna Muenke Popick & Fiona O'Donnell-McCarthy
Department of Psychology, Stanford University,
Jordan Hall, Stanford, CA 94305

Abstract

Although number words are common in everyday speech, for most children, learning these words is an arduous, drawn out process. Here we present a formal, computational analysis of number learning that suggests that the unhelpful structure of the linguistic input available to children may be a large contributor to this delay, and that manipulating this structure should greatly facilitate learning. A training-experiment with three-year olds confirms these predictions, demonstrating that significant, rapid gains in numerical understanding and competence are possible given appropriately structured training. At the same time, the experiment illustrates how little benefit children derive from the usual training that parents and educators provide. Given the efficacy of our intervention, the ease with which it can be adopted by parents, and the large body of research showing how strongly early numerical ability predicts later educational outcomes, this simple discovery could have potentially far-reaching import.

Keywords: Numerical Cognition, Learning Theory, Mathematical Modeling, Language, Learning

Introduction

Given the importance of numeracy to modern society, and the tortuous process of number learning experienced by many children, improving our understanding of how numbers are learned, and devising formal methods for improving this process, may produce numerous benefits for both individuals and societies. While number words are highly frequent in languages like English, appearing regularly in child-directed speech, children's acquisition of them is slow and labored (Wynn, 1992). Ask a three-year old for "3 balls," and they are likely to give you a handful instead, having treated '3,' rather indiscriminately, like 'some' (Wynn, 1990). This behavior does not stem from an inability to recognize differences between set-sizes: even 6-month-olds are able to discriminate between large set-sizes if the ratio is at least 2:1 (Xu, 2003; Xu, Spelke & Goddard, 2005; Lipton & Spelke, 2004) and this discriminability ratio becomes more fine-tuned over time (Wynn, 1998; Feigenson, Dahaene & Spelke, 2004; Van de Walle, Carey & Prevor, 2000) Children's difficulties with number are thus unlikely to be due to problems with detecting differences in quantity (Mix, Huttenlocher & Levin, 2002). Yet nor do they stem from an inability to grasp the relationship between language and quantity: one- and two-year-olds grasp that number words relate to quantities (Bloom & Wynn, 1997) and are often quite adept at reciting the count sequence (Fuson, 1988). The puzzle, then, is why children – who clearly both recognize number words as quantity designators and discriminate between set-sizes – go through an extended phase where they fail to understand how specific words match to specific quantities (Brannon & Van de Walle, 2001).

An ordinary child learning about number certainly will not suffer from any lack of exposure to count-relevant auditory and visual stimuli: count words and plural-sets are everywhere abundant. However, learning to discriminate which words match with which sets is not an insignificant problem: it involves 1) abstracting representations of specific set-sizes from the variable objects that make up any particular set, and then 2) mapping those representations on to specific number words. Here, we show how tightly coupled these processes are in learning (Gelman & Gallistel, 2004), and how they are effectively impeded by the way information is structured in English, and many other languages. We present a formal analysis and series of simulations that illustrate the problem and suggest a means of correcting it. In a training experiment, we then put this analysis to the test, contrasting the performance gains of children after typical number training – in which information was presented as usual – with that of children after restructured number training – in which the sequencing of linguistic information was manipulated to make it more conducive to learning and discrimination. The experiment reveals that when information is structured appropriately, 3-year olds rapidly improve their accuracy and consistency on not only trained number sets (2,4,6) but also on untrained sets (3,5,7). The improvement of the children following our intervention is particularly remarkable given that other recent training studies with older children have failed to find improvement even for trained numbers (Huang, Spelke & Snedecker, 2010) a finding replicated by the children in our 'typically structured' training condition.

Figure 1. An illustration of the challenge presented by number learning: there are nine objects: **one** red ball, **two** hats, **three** balls and **four** bears; there are **more** bears than balls or hats, **less** hats than balls, and **more** balls and hats than bears. Somehow, a child must discern the cues that discriminate between appropriate and inappropriate usage of each word.

Information Structure in Learning

One problem that a child learning number words must overcome is that she will never encounter numerical sets independently: she may encounter three apples, or three bears, but she will never encounter a "set of three" on its own (Wittgenstein, 1953). To further complicate matters, it

is virtually impossible to ascertain the meaning of a given number word from a single encounter (Fig. 1). For example, for a child faced with two apples and three oranges, the cues to the words "2" and "less" and "3" and "more" will initially be identical. This creates a discrimination problem: over time, a child must learn to discriminate which features appropriately match a given word in a given context.

In both natural and computational models of learning, this kind of discrimination is usually achieved by adjusting the degree to which various features in the environment are valued in predicting a relationship: highlighting those features which are most informative, and downgrading those which are not (Rosenblatt, 1959; Rescorla & Wagner, 1972; Gallistel, 2003). This suggests that over the course of number learning, the value of the features that successfully predict number words should increase, while the value of those that prompt erroneous expectations should correspondingly decrease. This process will produce competition for value between features, enabling the most reliable feature(s) to win out.

Given that in number learning, the best predictor of a given number word is set-size, the 'goal' of number learning is one of homing in on, and valuing, set-size over other competing features. So long as a given set-size – say, three – is the most reliable predictor of "3" in the environment, this goal will naturally be met as a result of the process of competitive reinforcement learning (Rosenblatt, 1959; Rescorla & Wagner, 1972; Gallistel, 2003), which will allow a child to discover and form a strong association between set-size three and the word "3," while simultaneously weakening any spurious associations to "3". With the correct association in place (and with ever-reducing interference from competitors), a child will then be able to accurately use and comprehend "three" (Fig. 2).

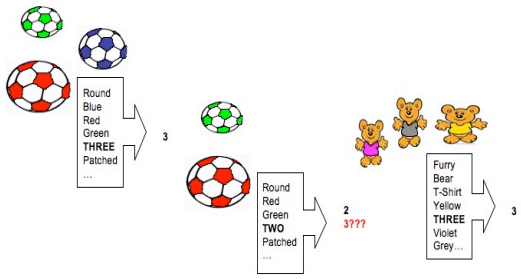


Figure 2. Here, we illustrate how the number *three* is learned over time. Learning is facilitated both by positive evidence (hearing the word "3" after seeing sets of three) and negative evidence (not hearing "3" when it is expected). Initially, several cues potentially predict "3," including uninformative features like round and red (**Left Panel**). However, these uninformative features will later erroneously cause "3" to be expected (**Center**). Because these unhelpful cues will result in prediction-error when, e.g., "2" is heard instead, they will lose value as cues to "3," both in this instance, and in any other cases where they erroneously predict "3." Further, because discrimination learning is competitive, they will lose associative value to more reliably informative cues. As set-size three continues to accrue positive evidence (**Right),** it will steadily gain value with respect to the initial set of cues. Provided that the relationship between the labels and the set-sizes is reliable, set-size three will eventually be learned as the meaning of "3" (see Ramscar et al., 2010).

However, given that this kind of learning is driven by prediction, the temporal structure of information will play a critical role in whether or not competitive learning actually occurs. Indeed, the effects of competitive learning can be isolated by comparing learning when complex (multi-feature) stimuli predict a series of discrete classes, to the inverse process (Ramscar et al., 2010). As Figure 2 shows, learning to predict a discrete Label – such as "2" or "3" – from a complex set of Features (FL-learning) allows for competitive learning amongst features, causing value to shift from features that produce more error to those that produce less. However, when this arrangement is temporally reversed, and the process becomes one of learning to predict a complex set of Features from a discrete Label (LF-learning), competition between cues cannot occur, since the label is the only cue present (value cannot transfer to other cues when there are none). Although these two processes appear similar, the differences in their temporal sequencing result in their having markedly different information structures, which produce very different patterns of learning (Ramscar et al., 2010). Color, another aspect of vocabulary that children master only after a noticeable delay (Darwin, 1877), offers an apt illustration of this.

Children's pattern of delay in learning colors words bears a striking resemblance to the pattern observed in number learning. Although color words appear in children's vocabularies from a very young age, sighted children's early use of them is comparable to that of blind children: i.e., they can produce them in familiar contexts ("yellow banana"), but cannot pick out novel objects by color, or reliably apply color words in unfamiliar contexts (Landau & Gleitman, 1985). Here again, children do not appear to grasp how specific words match to specific hues. Colors and numbers share several notable characteristics that may help explain the common pattern. First, like numbers, colors are properties of the environment, and cannot be encountered independently. Second, as with set-sizes, many different shades of color are present in any given context (Fig. 1). This means that in order to learn to map colors to their labels, a child must somehow discriminate the range of hues that best predict a specific color label from an environment in which color is ubiquitous (Ramscar et al., 2010). Fortunately, the difficulty of this problem can be significantly reduced if a child is encouraged to localize mappings – e.g., by seeking to extract color matches from known objects. This situation will allow the environment to be sampled in way that is far more informative (Landau & Gleitman, 1985). (Unfortunately, the structure of many languages proves largely unhelpful to learners in this regard; Ramscar et al., 2010).

To understand why, consider a child learning about the relationship between the features of a ball and various color labels, as depicted in Fig. 3. There are two possible ways this process can be structured temporally: either the various Features of the ball can predict the color Label (Feature-to-Label-learning) or the color Label can predict the ball's Features (Label-to-Feature learning; Ramscar et al., 2010).

Critically, the results of learning from these information structures differs markedly, which has important consequences for other sequential processes, such as language (Ramscar et al., 2010; Ramscar, Yarlett & Dye, 2009).

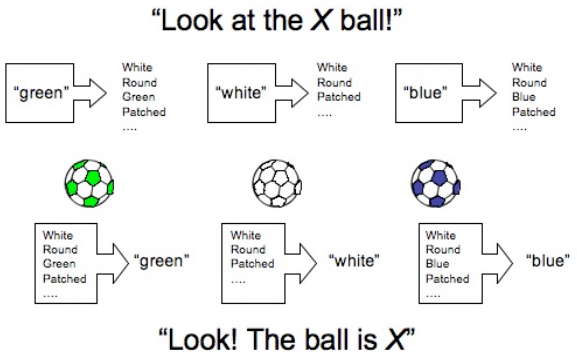


Figure 3. Learning can be dramatically affected by how information is presented to a learner in time (Elman, 1990; Ramscar et al., 2010). In this scenario, a child learns about the relationship between the features of a ball and various color labels. There are two ways this process can be structured temporally: either the child hears the color word used postnominally, which promotes **FL**-learning (the Features of the ball predict the color Label), or the child hears the color word used prenominally, which promotes **LF**-learning (the color Label predicts the ball's Features; Ramscar et al., 2010). Prior research on color learning indicates that *only* postnominal (FL) usage facilitates accurate category learning, whereas prenominal (LF) usage does not. Unfortunately, color words occur prenominally around 70% of the time in English (e.g., "the red ball;"), which may help explain English-speaking children's typically delayed pattern of acquisition (Rice, 1980).

Because children track events in their environment as speech unfolds (Tanenhaus et al., 1995; Kamide et al., 2003; Dahan & Tanenhaus, 2005; Fernald et al., 2006), the sequencing in an English sentence employing a postnominal construction ("Look! The ball is blue,") will present the feature information a child needs for color-label discrimination prior to the label that needs to be learned about. If the child has already learned "ball," her attention will be drawn to the ball before the word "blue" is heard. This means that postnominal constructions will typically result in FL-learning. However, this will not hold true for a prenominal construction ("Look at the blue ball"), where the label to be learned is heard prior to the known label. In this case, LF-learning will result.

The outcome of these two processes differs dramatically. In FL-learning, all of the features of the ball will be highlighted as potential cues to "blue," and with experience, the unreliable features (such as shape, size and texture) will lose value to the most reliable feature (color). Over time, and learning trials, this will result in representations in which features are valued relative to their informativity – that is, how well they predict the relevant label, given both positive and negative evidence. This will allow a child to learn the meanings of each of the color labels perfectly. By contrast, in LF-learning, there is no opportunity for competitive learning amongst features, and as a consequence, the child will develop a simple, probabilistic representation of the relationship between the label and object features, which captures co-occurrence information rather than informativity. This representation will impair category discrimination, since the overlapping, unreliable features will never fully be 'unlearned' (for a review, see Ramscar et al., 2010). Consistent with this, a prior study found that training with postnominal constructions significantly improved the accuracy and consistency of two-year olds' color word application, whereas a similar schedule of prenominal training had no effect on performance at all (Ramscar et al., 2010).

This raises the question of whether information structure plays a similar role in the acquisition of number words. Number words in English – and many other languages – are far more likely to occur in a prenominal position (e.g., "those three chairs"), than in a postnominal position (e.g., "those chairs, all three of them"). If our analysis is correct, hearing a number word postnominally should facilitate competitive discrimination learning, as the child discriminates what it is about, say, 'those chairs,' that predicts the word 'three.' However, so long as number words occur prenominally, the child will have no way of isolating the semantic cues (set-sizes) that best match number words.

Simulating Number Learning

To formally illustrate the problems involved in learning number, we conducted three sets of simulations. The first simulated the effects of prenominal and postnominal presentation on number learning; the second examined the effects that the peculiar information structure of number sets has on number learning; and the third integrated these factors, to examine predicted learning outcomes.

The effects of learning were simulated using the Rescorla-Wagner model (1972), a widely used learning rule that has been applied to numerous learning effects in animals and humans, and for which there is strong neurobiological evidence (Waelti, Dickinson & Schultz, 2001; Schultz, 2006; Niv & Schoenbaum, 2008). While it cannot account for all the phenomena observed in associative learning, the model provides an accessible formalization of the basic principles of error-driven learning, and is sufficiently detailed to allow a straightforward testing of the analysis we present here.

Simulation 1 modeled the learning of the association of sets of 2, 4 and 6 objects (with color, shape and size dimensions) with the labels "2," "4" & "6." Two simulations were implemented, one in which the sets and object features served as cues to the number labels (Feature-to-Label, FL), and one in which the number labels served as cues to the sets of objects and their features (Label-to-Feature, LF). Figure 4 illustrates why learning where object Features predict Labels (FL-learning) should result in far better learning of number words than when Labels predict Features (LF-learning).

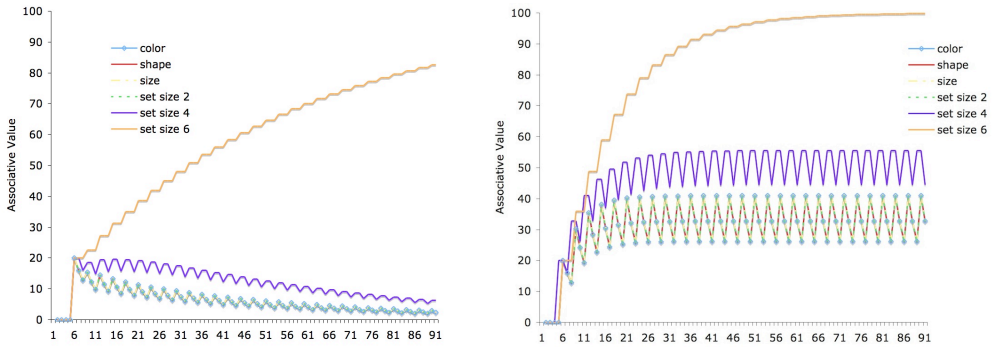


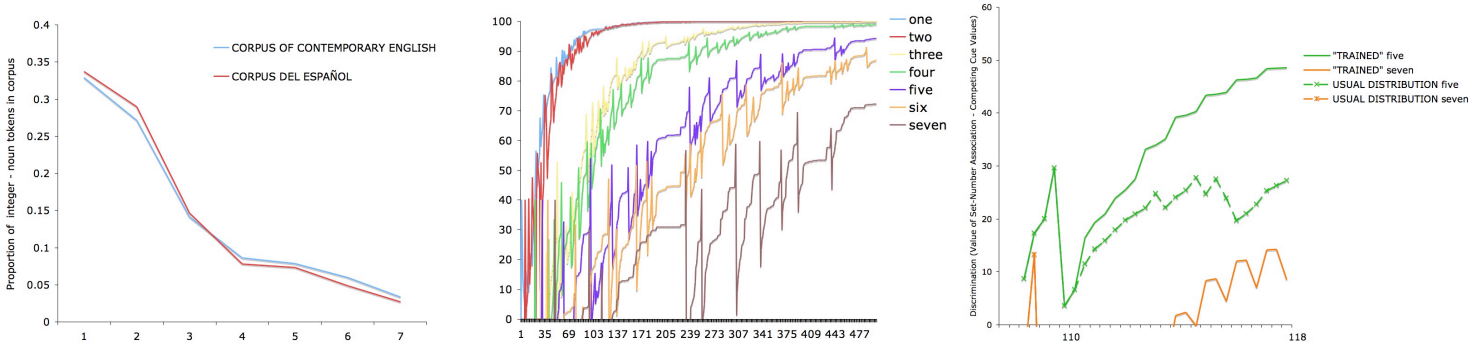
Figure 4. Simulations of number learning in which object **F**eatures predict **L**abels (**FL**-learning; left panel), and in which **L**abels predict **F**eatures (**LF**-learning; right panel). The models learned to associate sets of two, four and six objects to the labels "2," "4" and "6." In addition to number, each object set had size, shape and color cues that competed as cues with set-size as predictors of number words. These graphs depict the value of mappings between the object features, set-sizes and the label "6" learned in each simulation. As can be seen, FL-learning resulted in considerably greater discrimination of the appropriate cue-label mapping (set-size six to "6") than LF-learning, where competing activations continued to cause interference.

In Simulation 1, all set-sizes and numbers were experienced with equal frequency. However, it is unlikely that this is the case in real life. To get an estimate of the distribution of different set-sizes children might actually be expected to encounter and learn from, we examined the spoken distribution of number words in two languages (English and Spanish), taking frequency of mention as an index of the relevance of various set-sizes in children's lives. Both languages revealed the same distributional pattern, with the rank frequency of number words decreasing by quantity, following an inverse power function (Benford, 1938): "one" was the most frequent number word, followed by "two," "three," and so on (Fig. 5). This means the larger the set, the less frequently it is experienced.

At the same time, cue-competition should increase steadily with set-size: while the cue to set-size one is present in every set, the cues to "two" are only in every set greater than one, the cues to "three" are only in every set greater than two, and so on. Greater cue competition will demand a greater error signal to successfully resolve itself. However, since the extra competitors to larger sets will themselves be ever larger and less-frequent, larger sets will generate less and less of the error that makes discrimination learning possible. This means that confusability – and error – are unequally distributed in number sets, and leads to a intriguing situation with regards learning: as set size increases, the problem of discrimination gets successively harder, requiring increasing amounts of information to facilitate learning, just as the information available to the learner is shrinking.

To examine how the distribution of error in different sets might interact with the environmental relevance of different set-sizes, Simulation 2 was trained on sets in proportion to their spoken frequency. Specifically, the simulation modeled how the features of sets of 1-7 objects were associated with the labels 1 to 7. The simulation assumed that learners can discriminate objects from one another, and can contextually discriminate objects that are part of larger sets from objects that are not part of a larger set (i.e., that a learner can use context to discriminate a person standing alone from the same person standing with someone else). These elementary assumptions were reflected in the cue structure available for learning.

As Figure 6 illustrates, while learning to discriminate sets 1, 2 and then 3 and 4 was relatively straightforward, discriminating sets 5 and 6 required markedly more training, and discrimination of set size 7 remained poor, even after hundreds of training trials.



Figures 5 (Left Panel). The proportional frequency with which the numbers 1-7 are used to describe nouns in spoken English and Spanish (r=.999) (Davies, 2009; 2010). The distribution of number words by size and frequency follows an inverse power function: "one" is the most frequent number word, followed by "two," "three," etc. In the simulations, frequency of mention (the frequency of number word – noun sequences in a corpus) was used to estimate the relevance of different set-sizes in a learner's environment.

Figure 6 (Middle Panel). Learning to discriminate between set-sizes 1-7 after training on sets 1 - 20 according to their spoken frequency in English and Spanish (Davies, 2009; 2010). Sets 1-4 are discriminated straightforwardly, 5 and 6 require markedly more training, and 7 is discriminated only very slowly.

Figure 7 (Right Panel) . An illustration of how learning set-size could be impacted by training. In this simulation, training reflected the usual distribution of set sizes as suggested by English spoken frequency (Davies, 2009) for 110 trials, after which training either continued to reflect this distribution (the dashed lines represent the average of 5 such simulations) or else simulated exposure to six groups of 2, 4 or 6 objects learned FL (solid lines). The model trained on 2, 4 and 6 showed a marked improvement in its discrimination of 5 (solid green) and 7 (solid orange) despite not being trained on those items (Ramscar et al. 2010). This change was a result of the increase in the amount of error generated by 4 and 6, which in turn acted to increase the discriminability of 5 and 7.

The pattern of learning this produces appears to conform neither to the incremental nature of number sets, nor to Weber's law, which states that fixed levels of discrimination should occur between proportional set-sizes (i.e., 1:2 and 5:10 should be equally discriminable). Given that the input to this simulation comprised straightforward assumptions about the representation of sets and the environment in which they are learned, this result is striking. There has been much debate in the number literature over whether the differences in the way that smaller and larger sets are processed is evidence for a specific, capacity-limited system for representing small sets (Revkin et al., 2008), or whether the representation of smaller and larger sets is continuous (Cordes et al., 2001).This simulation reveals how, once the environment and the representational requirements of sets are taken into consideration, a continuous system for learning, representing and discriminating set-sizes can give rise to effective discontinuities in processing. This finding suggests one way in which these opposing perspectives might be formally reconciled, while leaving open the question of whether these differences are purely the result of learning (Cordes et al., 2001), or whether these constraints may begin to account for why the discrimination of smaller sets is hard-wired (Revkin et al., 2008).

Finally, Simulation 3 extended Simulation 1 by adding representations of size and shape to the sets of objects, as competing cues. Like Simulation 2, however, this simulation examined the effect that FL-training would have on a model previously trained on a more 'natural' distribution of sets: i.e., that observed in English and Spanish. The simulation was trained for 110 trials on the usual distribution with which numerical terms are related to sets in spoken English and Spanish (i.e.,the frequency with which number words are used to describe sets of nouns in each language) and then for 18 trials on a repeated pattern of sets of 2, 4 and 6 objects, to replicate the FL-training blocks of the three-year olds in our experiment. Figure 7 shows how six FL-training blocks of even sets (2,4,6) actually improved discrimination of *untrained*, odd sets (5,7). (This is a natural consequence of error-driven learning, see Ramscar et al., 2010 for a review).

As part of Simulation 3, we also ran five further simulations in which the last 18 trials were trained on the usual distribution of numerical terms in spoken English, and an average of the associative strengths learned between the cues and labels in these trials was taken for the purposes of comparing learning under "normal conditions" with the training simulation (see broken lines, in Fig. 7).

Training Experiment

We have described how a child might learn number words. The question is, *do* children learn in this way? Can manipulating the typical information structure of words in English – by teaching numbers in postnominal contexts – improve children's understanding of number?

Participants

Participants were 56 typically developing, monolingual English learners from 30 to 40 months old (M = 35.7 months, 30 females, 26 males) recruited from the Stanford area. Testing was conducted by an experimenter blind to the hypotheses.

Procedure and design

To test our predictions, we asked 56 children, aged 30 to 40 months, to identify twelve sets of objects on the basis of the numerosity. This established a baseline of competence for the numbers 2 through 7. Half of the pre-test questions were phrased pre-nominally ("Look! Can you show me four hearts?"), and half postnominally ("Look! Hearts. Can you show me four?").

Children were then randomly assigned to two training groups. In both conditions, children learned about the numbers 2, 4 and 6, with six familiar objects, which differed both in type and arrangement of presentation from those used in testing. The sets and labels employed in training were identical across conditions, with the critical distinction that the order of presentation was reversed. In the Feature-to-Label (FL) condition, a picture of the object set was shown first, and then the label was provided after the picture was shown ("What can you see? Balls. There are two"). By contrast, in the Label-to-Feature (LF) condition, the experimenter stated the number while the children looked at a blank page ("What can you see? There are two balls"), and immediately flipped to a picture depicting the object set as it was named. Thus, in the FL condition, children saw the object set and then heard the number label presented postnominally, while in the LF condition, children heard the number label presented prenominally, then saw the object set.

Children in each condition were then given a post-test identical to the pre-test.

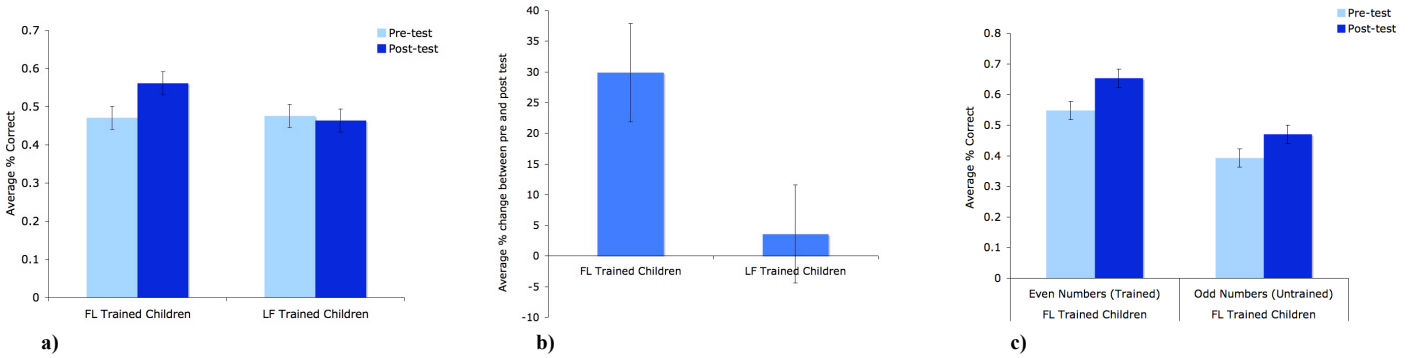


Figure 8. Performance in the identical pre-and post training-tests in the two groups of children (a) and average change in performance between the pre-and post-tests in the two groups (b). Graph (c) shows performance in the trained (even) and untrained (odd) pre-and post training-tests in the FL-trained children Because the untrained numbers were always tested together—separately from the trained numbers—the improvement on these items cannot be a result of children's improved performance on the trained items. (Error bars are SEM).

Results

Children's performance in these tests overwhelmingly supported our predictions about how the structure of information in training would affect children's ability to appropriately match set-sizes to their corresponding numerical labels. While there were no significant differences between the groups on pre-test performance (FL-trained M=47% correct; LF-trained (M=48% correct), the FL-trained children showed a marked improvement in the post-test (M=56%), whereas the LF-trained children (M=46%) did not (Figure 8a).

A 2 (item type: trained or untrained) x 2 (test type: pre versus post test) repeated measures ANOVA of children's performance (with training type—FL versus LF—as a between subjects measure) revealed that while overall performance had increased (there was a marginal effect of test type, $F(1,54)=3.399$, $p=0.07$), there were significant interactions between testing type and training-type ($F(1,54)=5.751$, $p<0.02$) and training-type and item type ($F(1,54)=4.44$, $p<0.04$), supporting the idea that FL-training was responsible for this improvement.

Planned tests revealed both that the FL-children's overall improvement in performance was significant (paired $t(27)=3.757$, $p<0.001$), and that this was true both on tests of the trained even numbers (pre-test M= 55%; post-test M=65%; $t(27)=2.447$, $p<0.025$) and the untrained odd numbers (pre-test M= 39%; post-test M=47%; $t(27)=2.555$, $p<0.01$); see Fig. 8c. LF-trained children's performance showed no change on either the trained (even) or untrained (odd) number tests (all tests $p>.3$). Overall, the FL-trained children performed 30% better on the post-test than the pre-test, whereas the change in the LF trained children was just 4% (unpaired $t(54)=2.242$, $p<0.05$); see Fig. 8b.

The different effects of training were further underlined by analyses of the consistency of the children's responses: First, the rate at which the LF-trained children provided consistent responses to tests of the same set-label mapping in the post-test (M=27%) was unchanged from the pre-test (M=28%), whereas the FL-trained children's post-test consistency again improved considerably (pre-test consistency M=30%, post-test M=38%), $t(27)=1.948$, $p<0.05$) (see Fig 8c); Second, FL-trained children's average performance improved across all of the items ($t(6)=2.824$, $p<0.05$), whereas the LF-trained children's average improved only for 3 and 6, and actually decreased slightly for 2, 4, 5 and 7 (this effect was not significant, $p>.4$).

Discussion

These data reveal that children as young as 2 ½ have begun to acquire an understanding of number words, and that this can be given a boost when the information structure in training supports competitive discrimination learning. FL-trained children, who saw the sets of objects before hearing labels presented postnominally, were significantly better both in terms of the accuracy and consistency of their reponses, both as compared to baseline measures, and in terms of their performance gains over LF-trained children. The performance of our FL-subjects was particularly remarkable given that longitudinal studies of 2 and 3-year-olds have demonstrated that improvements of this magnitude usually take place over months (Wynn, 1992), and not, as in our experiment, over half an hour.

Consistently using postnominal phrasing in child-directed speech, and introducing the object set (visually) before labeling it, may dramatically shorten the time-course of number word acquisition. Since a growing body of research suggests that understanding counting is predicated on a basic understanding of number (Wynn, 1990; Fuson, 1988; Branon & Van de Walle, 2001) and that mastery of this kind of numerical aptitude at a young age dictates later learning outcomes (Booth & Sigler, 2008; Jordan et al., 2010; Clements & Sarama, 2007) employing such an intervention may have a long lasting impact on children's mathematical aptitude and advancement.

Acknowledgments

This material is based on work supported by the National Science Foundation under Grant Nos. 0547775 and 0624345 to MR. This paper has been abridged. For further details about the model & a full list of references, please see:

http://psych.stanford.edu/~michael/numbers.pdf