

The Enigma of Number: Why Children Find the Meanings of Even Small Number Words Hard to Learn and How We Can Help Them Do Better

Michael Ramscar*, Melody Dye, Hanna Muenke Popick, Fiona O'Donnell-McCarthy

Department of Psychology, Stanford University, Stanford, California, United States of America

Abstract

Although number words are common in everyday speech, learning their meanings is an arduous, drawn-out process for most children, and the source of this delay has long been the subject of inquiry. Children begin by identifying the few small numerosities that can be named without counting, and this has prompted further debate over whether there is a specific, capacity-limited system for representing these small sets, or whether smaller and larger sets are both represented by the same system. Here we present a formal, computational analysis of number learning that offers a possible solution to both puzzles. This analysis indicates that once the environment and the representational demands of the task of learning to identify sets are taken into consideration, a continuous system for learning, representing and discriminating set-sizes can give rise to effective discontinuities in processing. At the same time, our simulations illustrate how typical prenominal linguistic constructions (“there are three balls”) structure information in a way that is largely unhelpful for discrimination learning, while suggesting that postnominal constructions (“balls, there are three”) will facilitate such learning. A training-experiment with three-year olds confirms these predictions, demonstrating that rapid, significant gains in numerical understanding and competence are possible given appropriately structured postnominal input. Our simulations and results reveal how discrimination learning tunes children’s systems for representing small sets, and how its capacity-limits result naturally out of a mixture of the learning environment and the increasingly complex task of discriminating and representing ever-larger number sets. They also explain why children benefit so little from the training that parents and educators usually provide. Given the efficacy of our intervention, the ease with which it can be implemented, and the large body of research showing how early numerical ability predicts later educational outcomes, this simple discovery may have far-reaching consequences.

Citation: Ramscar M, Dye M, Popick HM, O'Donnell-McCarthy F (2011) The Enigma of Number: Why Children Find the Meanings of Even Small Number Words Hard to Learn and How We Can Help Them Do Better. PLoS ONE 6(7): e22501. doi:10.1371/journal.pone.0022501

Editor: Dorothy Bishop, University of Oxford, United Kingdom

Received: March 24, 2011; **Accepted:** June 28, 2011; **Published:** July 27, 2011

Copyright: © 2011 Ramscar et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This material is based upon work supported by the National Science Foundation under Grant No. 0547775 to MR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: michael.ramscar@gmail.com

Introduction

Although number words are highly frequent in languages like English, and appear regularly in child-directed speech, children’s acquisition of them is slow and labored [1]. 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” [2]. 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 [3]–[6] and this discriminability ratio becomes more fine-tuned over time [7]–[9]. Children’s difficulties with number are thus unlikely to be due to problems with detecting differences in quantity [10]. 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 [1], [11] and are often quite adept at reciting the count sequence [1], [12]. 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* [13].

An ordinary child learning about number certainly will not suffer from any lack of exposure to count-relevant auditory and visual stimuli: count words are highly frequent and sets of items are everywhere. 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 [14] 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. Further, our simulations offer a solution to a puzzle relating to the nature of numerical knowledge: while most English speaking children will eventually learn to recognize and name sets of items in the small number range 1–4 without relying on counting [15], [16], in most cases, this ability does not reliably develop much beyond these values [17]. In our model, this pattern emerges naturally as a result of the discriminatory requirements of number learning, and the characteristics of the environment in which children learn numbers words.

A training experiment then puts this analysis of number learning 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 [18], a finding replicated by the children in our ‘typically structured’ training condition.

The Role of Information in Learning

In what follows, we describe and model the problem of learning number sets in learning and information theoretic terms. Given that straightforward applications of this approach are rare in language research, it is helpful to provide a basic outline of learning theory at the outset, particularly since contemporary models of learning represent a significant departure from the classic stimulus-response paradigm and do not share many of its limitations [19], [20].

Importantly, learning is no longer conceived of as simply a running tally of rewards and punishments; nor is it thought to be a process of accumulating simple associations between cues and outcomes in isolation. Instead, learning is best understood as a process that has evolved to help a learner better predict events in the world around her by weighing and assessing the **informativity** of cues for predicting relevant outcomes. In a similar vein, learning is no longer conceived of as simply a series of stimulus→response associations. Rather, it is understood as a process in which all the information available to a learner – both from the environment, and prior experience – is brought to bear on the task of predicting an outcome. Learning models describe the way that this information is sampled and processed for the purpose of better predicting events in the environment [21].

In line with this, experimental work in animal learning has demonstrated that when learning the predictive relationship between a given cue and a given outcome, animals do not simply chart how often cues predict certain outcomes, they also track how often cues *fail* to predict potential outcomes. The engine that

drives learning is not positive reinforcement, but surprise, or more formally, ‘prediction error’ [22], [23]. In learning models, prediction error is formalized as the discrepancy between the expected and actual outcomes a learner experiences, and learning is a process of incrementally updating a learner’s expectations in response to events [24].

These formalized learning rules have been shown to accurately predict the behavior of humans and animals across a wide variety of learning tasks, and to accurately reflect the firing patterns of mid-brain dopamine neurons [25]–[27]. When an event in a learner’s environment is incorrectly predicted, it provokes an error response [20]. This response is bidirectional: if an unexpected event occurs, dopaminergic activity spikes; if an expected event does *not* occur, activity dampens. More subtly, the strength of this spike – or dampening effect – is contingent on how poorly predicted the event was to begin with. Greater discrepancies between expectation and reality result in more error, and so more learning occurs; conversely, as discrepancies shrink, errors decrease in kind, and learning asymptotes [27]–[30].

Given the weight of behavioral and neurobiological support for this learning process, and the insight it offers into the way children learn of other verbal categories [21], we next consider whether it might help explain why children are so taxed by the challenge of acquiring an understanding of number.

Information Structure in English

The first 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 [31]. To further complicate matters, it is virtually impossible to ascertain the meaning of a given number word from a single encounter. 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 (Figure 1).

In many biological and computational models of learning, this kind of problem is solved by adjusting the degree to which various features in the environment are valued as cues to predicting a relevant outcome. This ‘adjustment process’ is competitive. Over the course of learning, features compete for predictive value, a contest which highlights reliably informative features, while downgrading or even eliminating uninformative features [21],



Figure 1. The challenges presented by number learning. This picture contains **nine** objects: **one** red ball, **two** hats, **three** balls and **four** bears; there are **more** bears than balls or hats, **fewer** 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. Unless one assumes that children’s *vocabulary* is innate, this problem will have to be solved even if children are granted some innate representation of number. That is, even if children have some internal concept of two, they still need to map the presence of two things in the environment to the word “2” and not to, say, “3,” which might be heard in the same context. doi:10.1371/journal.pone.0022501.g001

[23], [32], [33]. Characterized in these terms, number learning is a process of coming to value the appropriate set-size as the most reliable cue to a given number word, while at the same time discriminating it from other less reliable competitors (such as alternate set-sizes and other object features). The end goal is one of establishing which set-size best predicts which number word.

Notably, so long as a specific set-size is the most informative predictor of a number word in the learning environment, competitive discrimination learning ought to lead naturally to successful number learning [21], [23], [32], [33], allowing a child to discover and form a strong association between, say, 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 “3” (Figure 2).

However, the picture is somewhat more complicated than this suggests. Given that learning is driven by *prediction*, the temporal structure of information can play a critical role in whether or not competitive learning actually occurs. Indeed, the effects of competitive learning can be isolated by comparing learning in a situation where complex (multi-feature) stimuli predict a series of discrete classes, to its inverse [21]. As Figure 2 shows, learning to predict a discrete Label – such as “2” or “3” – from a complex set of Features (**FL**-learning) [21] 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 no other cues) [21], [23]. 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 [21]. This can be illustrated in relation to color, another aspect of vocabulary that children master only after a noticeable delay [34].

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: that is, 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 [35], [36]. 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 (Figure 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 [21], [37]. Fortunately, the difficulty of this problem can be significantly reduced if a child is encouraged to localize mappings – for example, by seeking to extract color matches from known objects. This situation allows the environment to be sampled in way that is far more informative [36]. Unfortunately, as we will show in a moment, the structure of many languages proves largely unhelpful to learners in this regard [21].

To understand why, consider a child learning about the relationship between the features of a ball and various color labels (Figure 3). As noted above, 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, **FL**) or the color **Label** can predict the ball’s **Features** (**Label-to-Feature** learning, **LF**) [21]. Because FL-sequencing produces competitive learning, whereas LF does not, the results of learning from these information structures differ markedly [21], [38]. However, as Figure 3 illustrates, which learning sequence results depends critically on how a child’s attention is directed in time, which, in turn, depends on whether the novel color label is introduced *before* or *after* the familiar noun.

Like adults, children track linguistically relevant events in their environment as speech unfolds in time [41]–[44], often directing their gaze at objects or object features as they are labeled in discourse. However, this kind of linguistically mediated visual attention requires that children actually know the meanings of labels. Because children learn the semantics of common nouns long before they learn those of common colors and numbers [45],

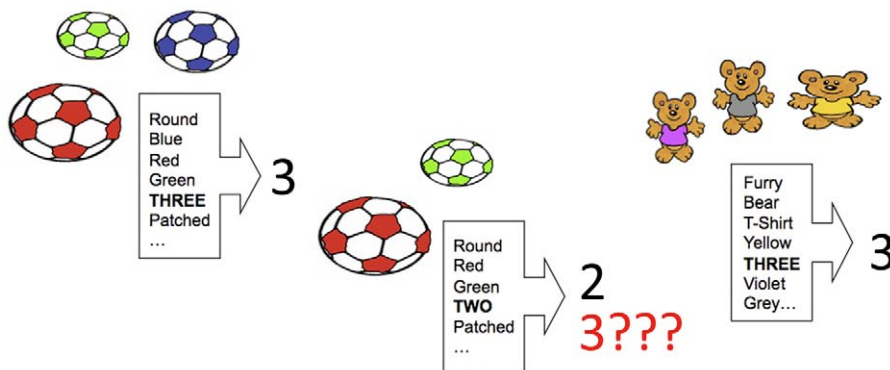


Figure 2. How the number three is learned over time. In competitive discrimination learning, positive evidence (reinforcement) increases associative value for cues, whereas negative evidence (prediction-error) correspondingly decreases value. In the **left** panel, each of the features present potentially predicts “3.” In the **center** panel, many of these unhelpful features will later erroneously cause “3” to be expected. Because these unhelpful cues will result in prediction-error when “2” is heard instead, they will lose value as cues to “3,” both in this instance, and in other cases where they erroneously predict a number word. Further, because discrimination learning is competitive, they will lose associative value to more reliably predictive cues (namely, set-size three). In the **right** panel, further positive evidence means that *three* continues to gain value with respect to the initial set of cues. As can be seen, learning is facilitated both by positive evidence – hearing the word “3” after seeing sets of *three* – and negative evidence – unlearning erroneous cues to “3,” like round and green. 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.”
doi:10.1371/journal.pone.0022501.g002

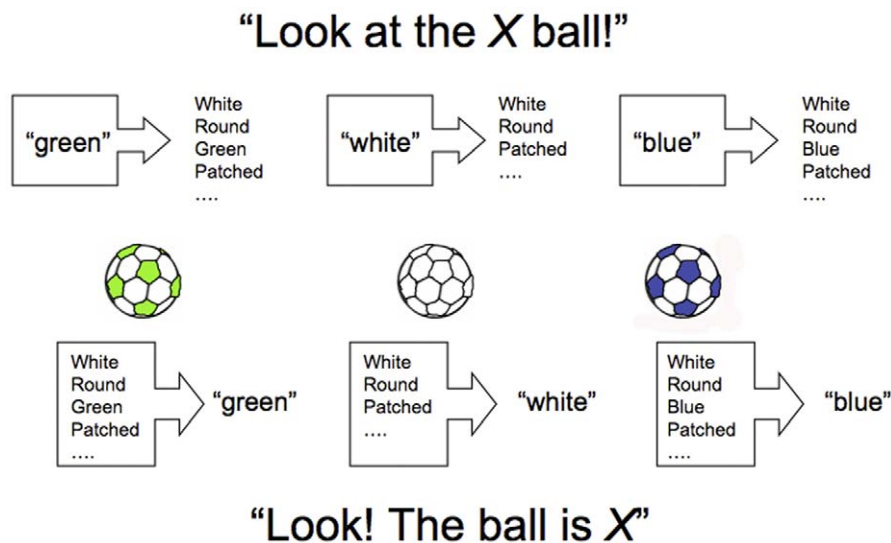


Figure 3. Linguistic sequence determines learning sequence. Learning can be dramatically affected by how information is presented to a learner in time [39]. Here, a child learns about the relationship between the features of a ball and various color labels. As illustrated, there are two possible ways this process can be structured temporally: either the child hears the color word used *postnominally*, which promotes **Feature-to-Label** learning (the **F**eatures of the ball predict the color **L**abel, **bottom** panel), or the child hears the color word used *preminally*, which promotes **Label-to-Feature** learning (the color **L**abel predicts the ball's **F**eatures, **top** panel) [21]. Prior research into category learning indicates that only FL-sequencing facilitates accurate category acquisition, whereas LF-sequencing does not [21,38,40]. doi:10.1371/journal.pone.0022501.g003

[46], a typical 2½ year old will readily direct her gaze toward a ball (or ball-like item) upon hearing the word “ball,” whereas a color word such as “blue” or “red” will not direct her visual attention in this way [47]. What this means, in practice, is that the sequence of events in an English sentence employing a *postnominal* construction (such as “Look! The ball is blue”) presents the information a child needs for color-label discrimination *prior* to the label that needs to be learned about, a sequence which supports FL-learning. However, the opposite is true for *prenominal* constructions (such as “Look at the blue ball”). Here, the color label is heard *prior* to the known label, which means that the child’s attention is not drawn to the ball until after she hears “blue.” Accordingly, prenominal presentation typically promotes LF-learning.

These two information structures can have dramatically different effects on learning. In **FL**-learning, all of the features of the ball are initially available as potential cues to “blue,” but with experience, unreliable features (such as shape, size and texture) lose value to the most reliable feature (color). This results in competitive learning, which produces *predictive* representations that value features relative to their informativity - that is, how well they predict the relevant label. Over time, this allows children to master the meanings of color labels [21]. By contrast, in **LF**-learning, competitive learning amongst features is not possible - as there is, in effect, just one feature - and as a consequence, a child will learn a simple, *probabilistic* representation of the relationship between the label and object features (specifically, the co-occurrence probability between the label and each feature, normalized by the probability of the label). Because overlapping, unreliable features will not be appropriately ‘unlearned,’ color discrimination will be poor. Consistent with this analysis, a prior study found that training with postnominal constructions (FL) significantly improved the accuracy and consistency of two-year olds’ color word application, whereas a similar schedule of prenominal training (LF) had no effect on performance at all [21].

Unfortunately for English-speaking children, however, color words are used preminally around 70% of the time in child-

directed speech [48], which may help explain why color acquisition is typically delayed [21], [35]. This also raises the question of whether information structure plays a similar role in the acquisition of number words. In English and many other languages, number words are far more likely to occur in a prenominal position (e.g., “those three chairs”), than in a postnominal position (e.g., “those chairs, the three of them”). If our analysis is correct, hearing a number word postnominally will facilitate competitive discrimination learning (helping a child discriminate what it is about, say, *those chairs* that predicts the word “three”), while instances in which number words occur preminally will be far less helpful to a child trying to learn to isolate the appropriate semantic cues to number words (i.e., set-sizes).

Of course, words are not the only cues that a child has to guide visual attention, and there may be alternate ‘routes’ to FL-learning, even when a prenominal expression is used. Research into joint attention has shown that children also make use of social cues such as gaze and gesture in learning to discriminate a word’s meaning [49]–[51]. For example, a parent might hand a child a handful of cookies before saying, “Here are *three* cookies,” or else point or look directly to a trio of cookies before mentioning their set-size. However, situations in which this kind of explicit instruction takes place are not representative of the majority of contexts in which children encounter number words [45]. Moreover, there is a great deal of variability in caregiver-child interaction during language learning: while some parents engage in frequent and sustained verbal interactions with their children, and explicitly label new objects, others only rarely communicate directly with their children, and do not engage in overt teaching behaviors [52].

To better isolate the effects of word order on number learning, we make a simplifying assumption here that all prenominal constructions support LF-learning, and all postnominal constructions support FL-learning. The analysis we present suggests that even learning in socially guided situations will benefit greatly from the information structure in postnominal constructions, and that postnominal ordering may be critical for learning in the majority

of contexts in which children naturally encounter number words in speech.

Analysis

To formally illustrate the problems involved in number learning, we conducted three sets of simulations. The first simulated the effects of prenominal (LF) and postnominal (FL) 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. All of the simulations employed the same reinforcement learning rule [23] and parameter values. Learning was simulated using the Rescorla-Wagner model [23], a widely used learning rule that has been applied to numerous learning effects in animals and humans, and which benefits from much neurobiological support [27]. While it cannot account for all the phenomena observed in associative discrimination learning, the model provides an accessible formalization of the basic principles of error-driven learning, and yet is sufficiently detailed to allow a straightforward testing of the analysis we present here.

Learning Model

The Rescorla-Wagner model simulates changes in the associative strengths between individual cues and an outcome as the result of discrete learning trials. If the presence of a cue or outcome X at time t is defined as $\text{present}(X, t)$, and its absence as $\text{absent}(X, t)$, then the predictive value V of a cue C_i to outcome O after a learning event at time $t+1$ can be specified as:

$$V_i^{t+1} = V_i^t + \Delta V_i^t \tag{1}$$

and changes (Δ) to the predictive value of V_i^t are defined as:

$$\Delta V_i^t = \begin{cases} 0 & \text{if } \text{ABSENT}(C_i, t) \\ \alpha_i \beta_1 (\lambda - \sum_{\text{PRESENT}(C_j, t)} V_j) & \text{if } \text{PRESENT}(C_j, t) \& \text{PRESENT}(O, t) \\ \alpha_i \beta_2 (0 - \sum_{\text{PRESENT}(C_j, t)} V_j) & \text{if } \text{PRESENT}(C_j, t) \& \text{ABSENT}(O, t) \end{cases} \tag{2}$$

Learning is thus determined by a discrepancy function in which λ is the total value of a predicted event O (the maximum amount of associative strength that the event can support) and V_j^t is the predictive value for O given the cues C_j present at time t .

When cues present on a trial are positively supported – i.e., where a predicted outcome occurs – the Rescorla-Wagner learning rule will produce a negatively accelerated learning curve (the result of events becoming better predicted, which reduces the discrepancy between what is expected and what is observed) and asymptotic learning over repeated trials (as events become fully predicted). When cues present on a trial produce **error** – i.e., when a predicted outcome fails to occur – λ (the value of the expected outcome) takes a value of zero because it didn't occur. In these cases, the discrepancy function will produce a negative value, resulting in a reduction in the associative strength between the erroneous cues and the absent outcome. Because error reduces the predictive value of cues, and because the total amount of value a given outcome can support is finite, this process causes cues to compete with one another for relevance in learning, leading to patterns of learning that usually differ greatly from those that would arise if learned values simply reflected correlations between cues and outcomes [19].

The rate at which learning takes place in the model (Δ) is determined by two factors: the overall learning rate β (where $0 \leq \beta \leq 1$), and the individual saliency of cues, α_i (where $0 \leq \alpha_i \leq 1$). Because we were interested in how learning affects the relative value of cues, α_i was set to 1, eliminating its influence on our simulations, while $\lambda = 100\%$ and $\beta_j = 0.2$. Thus the only free parameter was the learning rate, which we held constant in each implementation reported below.

Simulation 1: Sequencing Effects in Number Learning

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 “two,” “four” and “six.” 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 in

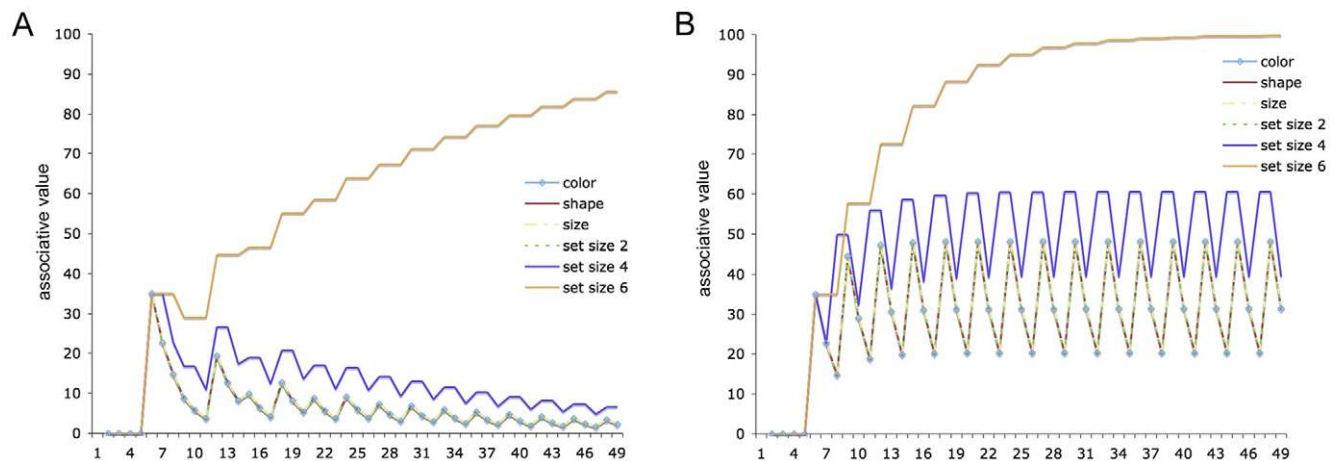


Figure 4. Simulation 1: The impact of temporal sequence on number learning. Panel A depicts a simulation of number learning in which object Features predict Labels (FL-learning), while Panel B depicts a simulation in which Labels predict Features (LF-learning). The models learned to associate sets of 2, 4 and 6 objects to the labels “two,” “four” and “six.” 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 “six” learned in each simulation. In FL-learning, uninformative cues are completely devalued as a result of cue competition, leading to enhanced discrimination.

doi:10.1371/journal.pone.0022501.g004

which object Features predict Labels (**FL**-learning) should produce far better learning of number words than when Labels predict Features (**LF**-learning). As can be seen, **FL**-learning result in considerably greater discrimination of the appropriate cue-label mapping (e.g., set-size 6 to “six”) than **LF**-learning, where competing activations continued to cause interference.

Simulation 2: Information Structure and Subitization

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 that sets of various sizes have 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: “one” was the most frequent number word, followed by “two,” “three,” and so on (Table 1). This suggests that the larger the set, the less frequently it is experienced (Figure 5).

At the same time, the discrimination problem a learner faces increases 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. This means that the confusability of sets – and the number of cues competing for value in each set – increases with set-size, which in turn increases the amount of error (i.e., training trials) that are required for the appropriate set-size cue to be successfully discriminated. However, given that 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. Because confusability – and error – are unequally distributed in number sets, this leads to an intriguing situation with regards learning: *as set-size increases, the problem of discrimination gets steadily 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 among different sets might interact with the environmental relevance of different set-sizes, Simulation 2 modeled how the features of sets of 1–7 objects were associated with the labels “one” to “seven” when sets were trained in proportion to their spoken frequency in English. The simulation made two assumptions: first, that learners can discriminate objects from one another, and second, that they can contextually discriminate objects that are part of larger sets from objects that are not part of a larger set (i.e., they 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. 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.

In recent years, there has been much debate in the number literature over whether the differences in the way that smaller and larger sets are processed – and, in particular, in the way that people’s ability to identify number sets without counting is limited to smaller numerosities – is evidence for a specific, capacity-limited system for representing small sets [56], or whether the representation of smaller and larger sets is continuous [57]. 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 (Figure 6A). This finding suggests one way in which these opposing perspectives might be formally reconciled, by showing how discrimination learning tunes the system for representing small sets, and how its capacity-limits result naturally out of a mixture of the learning environment and the increasingly complex task of discriminating and representing ever-larger number sets.

Table 1. The frequency of number words in English and Spanish.

CORPUS DEL ESPAÑOL	Tokens Per Million Words	COCA	Tokens Per Million Words
Uno	1079*	One	856
Dos	928.17	two	707
Tres	468.93	Three	368
Cuatro	248.99	Four	224
Cinco	234.11	five	204
Seis	155.44	Six	156
Siete	85.75	Seven	87

*estimate.

The table shows the spoken frequency counts of numbers 1–7 as they occur prenominally (e.g., “six hats”). The counts are taken from the 385 million word Corpus of Contemporary English (COCA) [53] and the 100 million word Corpus Del Español (CORDES) [54], respectively. (Note: The English-Spanish comparison is slightly complicated because “uno” is gendered in Spanish: it takes the form “una” with some nouns, and “una” is not used exclusively as a number word. The figure for “uno” presented here is a weighted estimate: $\frac{\text{number-word} + \text{noun sequences}}{\text{tokens of each number word in the corpus}}$.)

doi:10.1371/journal.pone.0022501.t001.

Simulation 3: Improving Set-Size Discrimination

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 (i.e., the frequency with which number words are used to describe sets of nouns, see Table 1), 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).

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 (broken lines, Figure 7).

For further discussion of these results, please consult the Postscript. R-files containing a re-implementation of the simula-

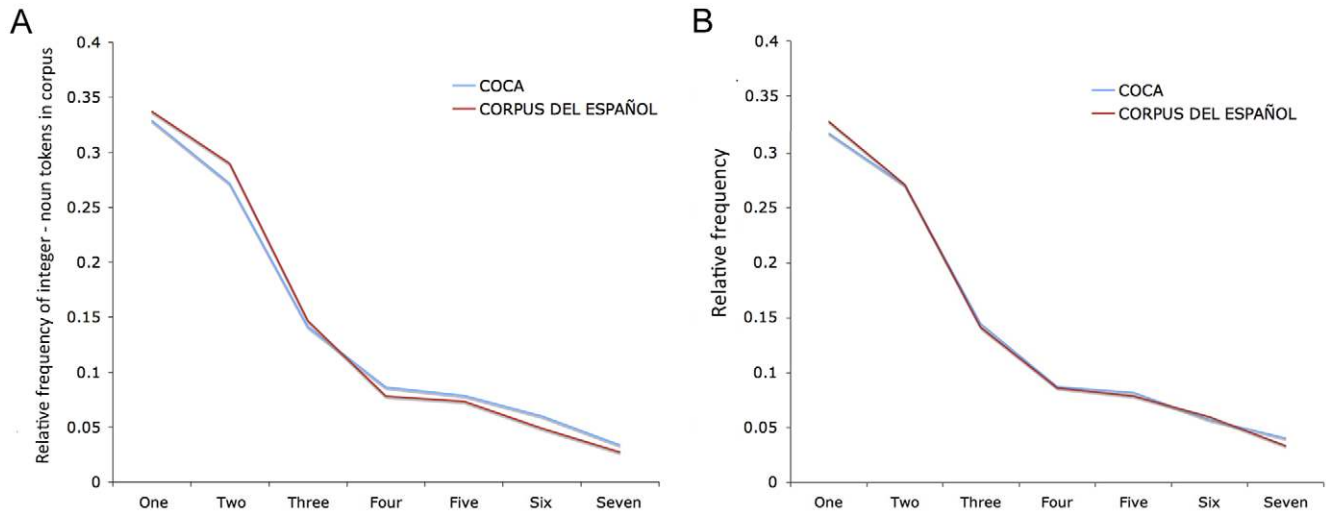


Figure 5. The distribution of number words follows an inverse power function in both English and Spanish. Panel A shows the relative frequency with which the numbers 1–7 are used to describe sets of nouns in spoken English and Spanish ($r = .999$) [53,54]. To ensure that the striking similarity in set mentions we found in the distribution of each language was not influenced by our weighted estimate of “uno,” we also examined the relationship between the frequency of number-word+noun sequences in English and the raw frequency counts for Spanish number words. Panel B is a graph of the relative frequency with which the numbers 1–7 are used to describe sets in spoken English [53] plotted against the relative frequency of the numbers 1–7 in the 100 million word Corpus Del Español [54]. Again, the same pattern and correlation ($r = .999$) was observed. These findings suggest that the distributions of number words in English and Spanish conform well to Benford’s law, which holds that lists of numbers from real-life sources of data will inevitably show an inverse power distribution [55]. We should note, however, that the probability distribution of numbers is somewhat more complex than this captures: because the decimal system – base ten – is employed for most everyday purposes, multiples of 10, 100, 1000, etc., tend to be used much more frequently than Benford’s law would predict, and similar, albeit smaller, peaks in usage frequency can be observed for multiples of five.
doi:10.1371/journal.pone.0022501.g005

tion, along with a Rescorla-Wagner (ndl) learning package, are available at <http://cran.r-project.org/web/packages/ndl/>.

Materials and Methods

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?

Ethics Statement

The experiments reported here were done in accordance with the Declaration of Helsinki. Additionally, they followed the ethical

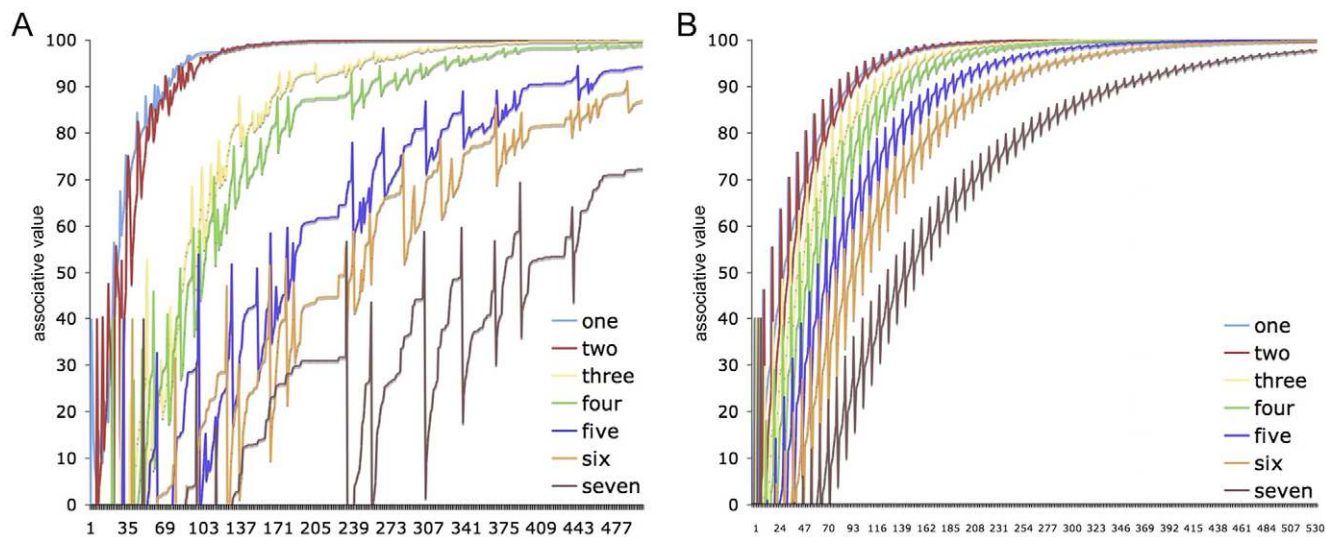


Figure 6. Simulation 2: Subitization is a product both of the learning environment and the representational requirements of discriminating larger sets. Panel A shows learning to discriminate between set-sizes 1–7 after training on sets 1–20 according to their spoken frequency in English and Spanish [53,54]. As can be seen, sets 1–4 are discriminated straightforwardly, 5 and 6 require markedly more training, and 7 is discriminated only very slowly. For comparison, Panel B shows learning to discriminate between set-sizes 1–7 after training in which the sets 1–8 were presented with equal frequency.
doi:10.1371/journal.pone.0022501.g006

requirements of the Stanford University institutional review board and complied with ethics guidelines set forth by the IRB recommendations. This study was specifically approved under IRB Protocol “The Acquisition, Representation, and Processing of Conceptual and Linguistic Knowledge” (ID: 14178, Number: 349, Group 2, 2008–2011). The parents of the participants were informed that participant data would be treated anonymously and that participants could terminate the experiment at any time without providing any reason. We received written informed consent from all parents before beginning the experiment.

Participants

Participants were 56 typically developing, monolingual English learners from 30 to 40 months old ($M = 35.7$ months) recruited from the Stanford area. 30 participants were female and 26 were male. Testing was conducted by experimenters who were blind to the hypotheses tested.

Procedure

The experiment comprised four stages: familiarization, Test-A (pre-test), training, and Test-B (post-test), with familiarization and testing following a three-alternative forced-choice procedure. All stages were identical for participants, except training, in which there were two conditions. Half of participants were randomly assigned to one training condition, and half to the other. There were 13 males and 15 females in each group, and age was matched between groups (FL-trained children, $M = 35.8$ months; LF-

trained children $M = 35.5$ months).

Familiarization. In the familiarization exercise, children were shown slides with three images of common objects (e.g., a dog, a tree, and a banana), and were asked to identify a particular object by pointing to it (“Can you show me the doggy?”). This exercise continued for ten slides, or until the child reliably pointed to the particular object that was requested every time, repeating slides if necessary. All participants reliably passed this stage of testing.

Pre-test. Test-A then evaluated the child’s ability to identify object sets on the basis of numerosity. The pre-test consisted of 12 slides, which each had a familiar object (heart, square, cloud, circle, sun or star) appearing in three different set-sizes (Figure 8). Each slide had sets of either 2, 4 and 6 objects, or sets of 3, 5, and 7 objects. Children were asked to identify one set per slide. Half the time, the child would be asked to identify a particular set with a prenominal phrased question (such as “Look! Can you show me *four* hearts?”); the other half, with a postnominal phrased question (such as “Look! Hearts. Can you show me *four*?”). Children were given identical positive feedback after each trial regardless of accuracy. Over the span of the test, each set-size was requested twice, and the position of the correct answer was counterbalanced across trials. This established a baseline of competence for the numbers 2 through 7.

The results of the pre-test revealed that children’s performance was above chance, suggesting that they had some experience matching number words to set sizes (chance in 3-choice procedure = 33.3%; FL-condition $M = 47\%$ correct; LF-condition $M = 48\%$ correct). However, their grasp of numerical concepts appeared less than expert (pre-test consistency, FL-condition $M = 30\%$; LF condition $M = 28\%$), indicating that their understanding of set sizes was still developing. (Notably, there was no significant difference in pre-test scores between children in the two training conditions.)

Training. In training, children were randomly assigned to one of two conditions. Across both conditions, children learned about the numbers 2, 4 and 6, with six familiar items (hats, bears, fish, cars, balls, and boats), in succession from small to large (e.g., 2 hats, 4 hats, 6 hats; then 2 balls, 4 balls, 6 balls; Figure 9). Notably, these items differed both in type and arrangement of presentation from those used in testing. On every training trial, each child would both see a slide (containing a depiction of one of the plural sets) and hear the appropriate label for that set-size and item. Because there were six different items, each child received six exemplars for each of the set-sizes, for a total of 18 training trials.

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 item 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 slide (“What can you see? There are *two* balls”), and then immediately flipped to a picture depicting the item set as it was named. Thus, in the **FL** condition, children saw the item set and then heard the number label (presented postnominally), while in the **LF** condition, children heard the number label (presented prenominal), and then saw the item set. Because training involved the explicit presentation of slides – to which the children’s attention would naturally be drawn – the sequencing of the presentation of slides and labels was intended to replicate the effect that word order would have on the way in which information became available to children in more naturalistic contexts [21].

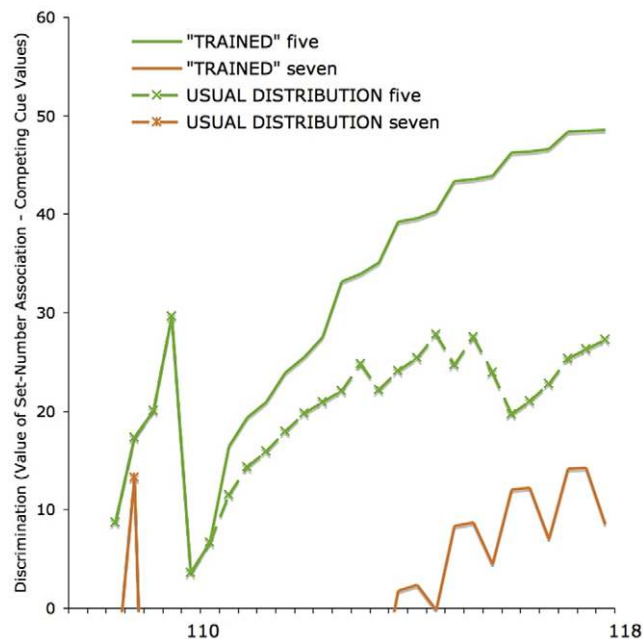


Figure 7. Simulation 3: Number learning can be facilitated by appropriate training. In this simulation, training reflected the usual distribution of set sizes as suggested by English spoken frequency [53] for 110 trials, after which training either continued to reflect this distribution (the dashed lines 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. 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. See postscript.

doi:10.1371/journal.pone.0022501.g007

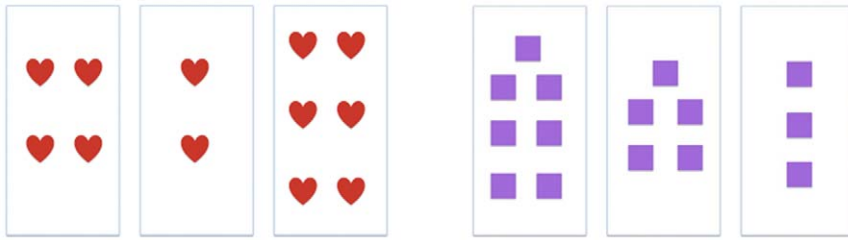


Figure 8. Examples of the arrays used in the pre- and post tests. Testing followed a three-choice procedure, putting chance performance at 33.3%.

doi:10.1371/journal.pone.0022501.g008

The only difference between the two conditions was temporal ordering in the presentation of the slides and verbal sequences.

Post-test. After training, children participated in Test-B, which was identical to Test-A. Given the inconsistent nature of children's number word usage and comprehension, this design allowed a measure both of the probability that children would match number word and set-size correctly on a given test, and also the probability that they would match number word and set-size correctly and *consistently* on consecutive tests. This allowed us to gauge differences in both the consistency and overall accuracy of children's number word knowledge after training.

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-condition $M = 47\%$ correct; LF-condition $M = 48\%$ correct), the FL-trained children showed a marked post-test improvement ($M = 56\%$, $stdev = 21$), whereas the LF-trained children ($M = 46\%$, $stdev = 21$) did not (Figure 10A).

For analysis, the pre- and post-test measures were transformed into z-scores. A 2 (item-type: trained or untrained) \times 2 (test-type: pre- versus post-test) repeated measures ANOVA of children's performance (with training-type – FL or LF – as a between subjects measure) confirmed the effect of training, revealing a main effect of test-type, $F(1,54) = 10.744$, $p = 0.01$, and significant interactions between the test-type and training-type ($F(1,54) = 4.554$, $p < 0.05$) and item-type and training-type ($F(1,54) = 6.288$, $p < 0.025$), supporting the idea that the FL-training was responsible for this improvement. There was also marginally significant effect of item-type, $F(1,54) = 3.839$, $p < 0.06$, which reflected the children's overall better performance on the even numbers – 2, 4 & 6 – than the odd numbers – 3, 5 & 7 – in both the pre- and post-tests. Given that 2 of the 3 smallest sets

tested were even, and 2 of the 3 largest sets were odd, this is consistent with the simulations, which suggest that larger set-sizes are harder to learn.

Planned tests revealed both that the FL-children's overall improvement in performance was significant (pre-test $M = 47\%$; post-test $M = 56\%$; $stdev = 21$; paired $t(27) = 3.885$, $p < 0.001$), and that this was true both on tests of the trained even numbers (pre-test $M = 55\%$; post-test $M = 65\%$; $stdev = 27$; $t(27) = 2.446$, $p < 0.025$) and the untrained odd numbers (pre-test $M = 39\%$; post-test $M = 47\%$; $stdev = 24$; $t(27) = 3.073$, $p < 0.01$); see Figure 11. LF-trained children's performance showed no change on either the trained (even) or untrained (odd) number tests (all tests $p > .3$). Overall, the average individual gain in performance between the pre-test and the post test (i.e., each child's post-test score/pre-test-score) for the FL-trained children was 30%, which was significantly greater than the change for the LF-trained children, which was just 4% ($stdev = 45$; unpaired $t(54) = 2.242$, $p < 0.05$); see Figure 10B.

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 significantly (pre-test consistency $M = 30\%$, post-test $M = 38\%$; $stdev = .28$; $t(27) = 1.499$, $p < 0.05$); Second, FL-trained children's average performance improved across all of the items (pre-test $M = 47\%$; post-test $M = 56\%$; $stdev = 16$; $t(6) = 2.825$, $p < 0.05$), 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

The results of this experiment reveal that children as young as 2 $\frac{1}{2}$ have begun to acquire an understanding of number words, and that this understanding can be given a significant boost when the



Figure 9. Examples of the items used in training. Sets of 2, 4 and 6 bears are shown. The training items were designed to maximize the predictive value of set-size as a cue by manipulating the error produced by competitor cues such as size, color and shape.

doi:10.1371/journal.pone.0022501.g009

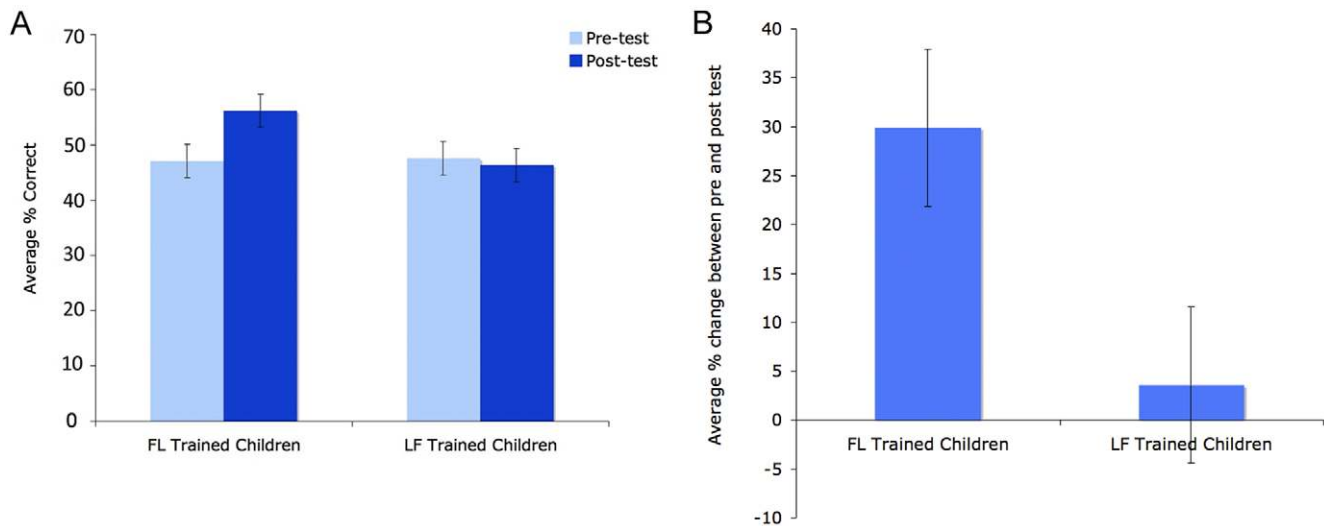


Figure 10. A comparison of performance between the FL- and LF-trained children. Panel **A** shows performance in the identical pre-and post-training tests in the two groups of children, while Panel **B** depicts the average individual change in performance between the pre-and post-tests for each group. Error bars are SEM.

doi:10.1371/journal.pone.0022501.g010

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 responses, both as compared to baseline measures established in the pre-test, and in terms of their overall 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 a time course of months [1], and not, as in our experiment, over the span of half an hour.

While the LF-training paradigm we employed here took particular care to control the information available to children, there is little reason to believe that a “more natural” prenominal

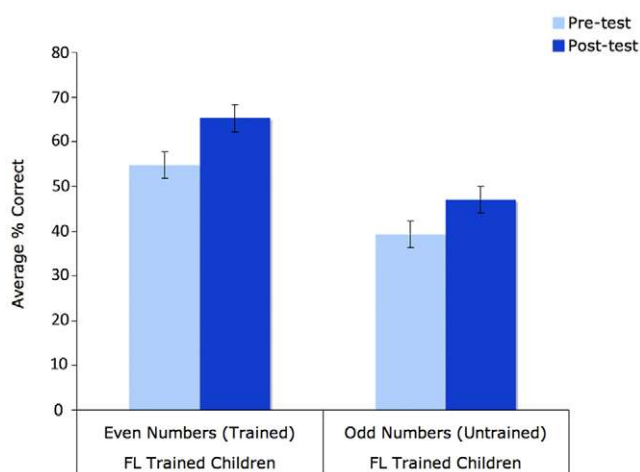


Figure 11. Performance in the trained (even) and untrained (odd) pre- and post training-tests in the FL-trained children. Error bars are SEM. Because the untrained numbers were always tested together – separately from the trained numbers – the improvement on these items cannot be a result of the children’s improved performance on the trained items.

doi:10.1371/journal.pone.0022501.g011

presentation – in which the items were available to the children as the sets were verbally enumerated – would have helped the LF-trained children. There are at least three reasons to think this: First, identical patterns of learning were observed in children given postnominal (FL) and prenominal (LF) training on color words while the objects were on display during the period they were described [21]. Second, the failure of the LF-trained children to benefit from their training is consistent with the findings of other studies that have employed more “natural” prenominal training, and in which children have similarly failed to benefit from training [18]. (Notably, previous training experiments with number words have all used prenominal instruction, likely because prenominal phrasing is highly preferred in English). Third, when older children and adults are presented with a similar task structure and asked to match *known* information – i.e. when learning isn’t an issue – they perform far better mapping from **L**abels-to-**F**eatures than **F**eatures-to-**L**abels, because labels themselves are far less complex than the things that they actually label, and make fewer demands on memory [58]. This suggests that there is a trade-off between communicative efficiency and learnability: once numbers have been successfully mastered, prenominal usage may have processing advantages [59], but until they are, prenominal usage can actually impede learning.

These findings underline the theoretical importance of information structure to considerations of human development, and suggest practical ways in which a better understanding of information structure can assist educators and parents who wish to speed or enhance the learning process. Consistently using postnominal phrasing in child-directed speech, and introducing object sets (visually) before labeling them, 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 [1], [2], [12], [14] and that mastery of this kind of numerical aptitude at a young age dictates later learning outcomes [60]–[63] interventions that utilize helpful information structures may have a long lasting impact on children’s mathematical aptitude and advancement. Additionally, other results indicate that manipulating the structure of information in the way we have outlined here can improve performance

across a range of perceptual domains and tasks, including visual and auditory category learning [21], [38], [64], and contextual rule learning [40].

Given the importance of numeracy to modern society, and the difficulty many children experience in grasping numerical concepts, improving our understanding of how numbers are learned, and devising formal methods for improving this process, can benefit both individuals and societies. The success of our training study shows that these methods need not be complex, and illustrates how formal analyses of information and learning can make principled predictions about how a specific intervention can rapidly (and beneficially) accelerate learning, revealing the benefits that formal analyses of learning can bring to education [19], [21], [40].

Postscript: The importance of being wrong

Historically, the computational and theoretical nature of formal discriminative learning models has been subject to widespread misunderstandings [19], [21], and it is likely that readers unfamiliar with discrimination learning may be left wondering why the children in the postnominal (FL) condition improved not only on trained (even) numbers, but also on untrained (odd) numbers, and why our simulation of learning (Figure 7) predicted this improvement.

To explain why this is so, it is important to understand that in discrimination learning, *expectations that are wrong shape learning more than expectations that are right*. This counter-intuitive principle underpins all formal models of learning, and it is perhaps most easily grasped in relation to animal learning. Consider a classic conditioning experiment in which a rat is subjected to a series of tones followed by mild shocks, and rapidly learns to respond fearfully to the tones. While intuitively, it might seem that learning is caused by the positive relationship between the tones and the shocks – which leads naturally to the idea that learning is a simple matter of relating stimuli to punishments and rewards – this is not the case. If tones that do *not* lead to expected shocks are added to the tone-shock pairings, rats' conditioned responses will weaken in direct proportion to the increased **background rate** of tones [65]. This makes good sense, because rats' responses hinge on how informative the tones are about the shocks, and if tones regularly occur without leading to shocks, then they are no longer obviously informative about shocks. Importantly, however, this requires that we accept that the rats are learning about what is *not happening*: in this instance, the rats' learning is driven by the non-occurrence of expected events, which result in neither punishment, nor reward.

Given that rats have no a priori knowledge about the relationship between the tones and the shocks, it may seem natural to wonder why it is that only the background rate of tones matters here. The answer, of course, is that it *isn't* only the background rate that matters. In principle, everything in the rat's experience and environment matters in predicting the shock. However, the rat's previous experience with other aspects of its environment – and what it has learned about *their* background rates – will have mitigated their influence on subsequent learning. For example, the degree to which the color of a rat's enclosure is learned as a cue to the shocks will be affected both by the rate at which that color has been previously experienced in situations that did not lead to shocks, and the rate at which that color is subsequently experienced in situations that do not lead to shocks. In all likelihood, the background rate of that feature will be sufficiently high that the rat will effectively ignore it as a potentially informative cue, even though it positively co-occurs with every shock. What is critical to note here, is that learning about the tones takes place within the greater context of the rat's experience with

other aspects of its environment, and against an extensive backdrop of possible cues and their associated background rates.

Thus, while learning theoretic models tend to focus on the tones – because they serve as novel, and hence, informative cues – it is important to keep in mind that this novelty is entirely relative. It only makes sense to say that a cue is perceived as novel if we assume that every other perceptual cue available is *not* as novel. Or, within the context of the tone-shock paradigm, it only makes sense to describe the tone as a novel perceptual cue if we assume that its novelty is computed in relation to all the other cues available to the rat [19]. In discrimination learning, knowledge of a positively informative relationship (e.g., between tones and shocks) results from the competitive elimination of less informative relationships. Since the latter inevitably outnumber the former, it follows that error-information is as valuable – if not more valuable – than information about successful predictions.

These ideas are essential to understanding our model of number learning, which assumes that children, like other animals, are sensitive to the value of information in their learning environment. The model assumes that in addition to tracking how often words and features are paired together (e.g., a cow is seen and “cow” is heard), children attend to how often a potential pairing does *not* occur (e.g., an object is seen and “cow” is not heard). Since children never encounter an example of a given numerosity unless a set of enumerated objects is also present, they are faced with the problem of discriminating the relevant dimension (set-size) from both the objects themselves and other properties of those objects. This is a hard problem. However, assuming that learning tracks how often pairings *don't* occur (e.g., four objects are seen and “three” is not heard), children should be able to home in on which dimension is the most informative about which number word, simply by learning to *ignore* other features based on their background rates.

How does this play out in our training experiment? Given that the pre-test scores of the children in our experiment were above chance, we can safely assume that they had already done some learning ‘in the wild’ about number words and the social contexts in which they could expect to hear them. At the same time, while it is clear that the children had some experience with numbers, and had heard them in a variety of contexts, their pre-test scores suggest they had not yet learned to fully discriminate the appropriate cues (i.e., specific set-size) to number words, or to discriminate various number words from one another (e.g., “three,” “four” and “five” might seem more or less interchangeable). In our model, this prior learning is represented by the first 110 trials in Simulation 3 (Figure 7).

In the training experiment, half of the children are then given FL-training in which – for example – they are shown a set of objects (say bears), and the experimenter asks, “What can you see? Bears. There are *four*.” As a result, each child learns about the relationship between the pictorial representation in front of her (the bears) and the number word (four). But what, precisely, does she learn about this relationship? When the experimenter says, “Bears – there are *four*” –

1. All of the features present (e.g., color, shape, set-size, etc) will be **reinforced** as cues to “four.”
2. Given that the child is playing a number game, in which number words have been generated from similar items, she will be expecting to hear other numbers as well – such as “three” and “five” – based on prior experience. This will generate **error** for all of the features that prompted those erroneous expectations, which will in turn *cause the features present to be unlearned as cues to other number words*. This means that not only

will set-size four be unlearned as a cue to “three” and “five,” but so will all the consistently unreliable features, such as color, shape, and so on.

Thus, while 3, 5 and 7 were not explicitly trained in the experiment, the prediction errors generated on 2, 4 and 6 trials will have helped children learn to discriminate the appropriate cues to the odd numbers – and increase the informativity of their representations of those numbers – *even though there was no positive reinforcement for 3, 5 and 7*. For example, while the actual cue value of set-size five cannot be affected by learning on a 4-trial (because it is not present), other potential predictors of set-size five will be present on that trial (e.g., color, shape, and so on), and their value as cues to “five” will be affected by this. The non-occurrence of “five” will cause the value of cues that erroneously predicted “five” to decrease, which will cause a *relative* increase in the value of set-size five as a predictor of “five.” As a result, the child will be able to better discriminate the cues to “five” (i.e., she will have learned about “five”), even though – or, in fact, because – “five” was not heard in this context.

Perhaps counterintuitively, this also means that for a child who has some experience of “four,” but has yet to fully discriminate the mapping between set-size four and “four,” trials in which “four” is presented will not be particularly informative about “four.” This is because the cue value of set-size four will not change relative to the other cues present on 4-trials, since they will all be similarly reinforced as predictors of “four.” On the other hand, a 4-trial will increase the **background rate** of all the cues that prompt erroneous expectation of *other number words*. Thus, for example, shape might be devalued relative to set-size five as a cue to “five,” and color might be devalued relative to set-size three in predicting “three” (and so on), and over time this process will help the child improve her discrimination of a system of numbers.

Given our comments on the systematic nature of discrimination learning, we hope it is clear that the real learning process in a child is far more complicated than the simplified version in our model. With that caveat in mind, however, the structure of our training was designed to increase the background rate of the erroneous cues relative to set-size, and thus – as a *result* of the error generated by the irrelevant cues – help our subjects more strongly associate the appropriate set-sizes with the appropriate labels. Because the unlearning of erroneous cues is an important, active part of discrimination learning, and because the structure of the task was designed to generate error over 3, 5 and 7, the design of this experiment provided a strong test of our learning hypothesis [66]. That the children did improve on 3, 5 and 7 provides support for this hypothesis, and illustrates the potential benefits that the application of discrimination learning models and analyses of information structure can bring to our understanding of children’s learning.

Acknowledgments

We are grateful to Brad Love, Harald Baayen, Adele Goldberg, Daniel Yarlett and Rich Shiffrin for discussion of these ideas and to Dorothy Bishop for her invaluable stewardship over this paper. Heartfelt thanks are offered to Robert Ryan and Christine Tran for heroic feats of data collection.

Author Contributions

Conceived and designed the experiments: MR. Performed the experiments: MD HMP FO-M. Analyzed the data: MR MD. Contributed reagents/materials/analysis tools: MR FO-M. Wrote the paper: MR MD HMP.

References

- Wynn K (1992) Children’s acquisition of the number words and the counting system. *Cognitive Psychology* 24: 220–251.
- Wynn K (1990) Children’s understanding of counting. *Cognition* 36: 155–193.
- Xu F, Spelke ES (2000) Large number discrimination in 6-month-old infants. *Cognition* 74: B1–B11.
- Xu F (2003) Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition* 89: B15–B25.
- Xu F, Spelke ES, Goddard S (2005) Number sense in human infants. *Developmental Science* 8: 88–101.
- Lipton JS, Spelke ES (2004) Discrimination of large and small numerosities by human infants. *Infancy* 5: 271–290.
- Wynn K (1998) Psychological foundations of number: Numerical competence in human infants. *Trends in Cognitive Science* 2: 296–303.
- Feigenson L, Dehaene S, Spelke ES (2004) Core systems of number. *Trends in Cognitive Science* 8: 307–314.
- Van de Walle GA, Carey S, Prevor M (2000) Bases for Object Individuation in Infancy: Evidence from Manual Search. *Journal of Cognitive Development* 1: 249–280.
- Mix KS, Huttenlocher J, Levin SC (2002) Multiple Cues for Quantification in Infancy: Is Number One of Them? *Psychological Bulletin* 128: 278–294.
- Bloom P, Wynn K (1997) Linguistic cues in the acquisition of number words. *Journal of Child Language* 24: 511–533.
- Fuson KC (1988) Children’s counting and concepts of number. New York: Springer-Verlag. 446 p.
- Brannon EM, Van de Walle GA (2001) The development of ordinal numerical competence in young children. *Cognitive Psychology* 43: 53–81.
- Gelman R, Gallistel CR (2004) Language and the origin of numerical concepts. *Science* 306: 441–443.
- Jevons WS (1871) The power of numerical discrimination. *Nature* 3: 363–372.
- Kaufman EL, Lord MW, Reese TW, Volkman J (1949) The discrimination of visual number. *American Journal of Psychology* 62: 498–525.
- Vetter P, Butterworth B, Bahrami B (2008) Modulating Attentional Load Affects Numerosity Estimation: Evidence against a Pre-Attentive Subitizing Mechanism. *PLoS ONE* 3: e3269. doi:10.1371/journal.pone.0003269.
- Huang Y, Spelke ES, Snedeker J (2010) When is ‘four’ far more than ‘three’? Children’s generalization of newly-acquired number words. *Psychological Science* 21: 600–606.
- Rescorla RA (1988) Pavlovian conditioning: It’s not what you think it is. *American Psychologist* 43: 151–160.
- Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology* 57: 87–115.
- Ramsar M, Yarlett D, Dye M, Denny K, Thorpe K (2010) The Effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science* 34: 909–957.
- Kamin LJ (1969) Predictability, surprise, attention, and conditioning. In Campbell B, Church R, eds. *Punishment and aversive behaviour*. New York: Appleton-Century-Crofts. pp 279–296.
- Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, eds. *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts. pp 64–99.
- Niv Y, Schoenbaum G (2008) Dialogues on prediction errors. *Trends in Cognitive Science* 12: 265–272.
- Montague PR, Dayan P, Sejnowski TJ (1996) A Framework for Mesencephalic Dopamine Systems based on Predictive Hebbian Learning. *Journal of Neuroscience* 16: 1936–1947.
- Schultz W (1998) Predictive reward signal of dopamine neurons. *Journal of Neurophysiology* 80: 1–27.
- Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412: 43–48.
- Schultz W, Dickinson A (2000) Neural coding of prediction errors. *Annual Review of Neuroscience* 23: 473–500.
- Schultz W, Dayan P, Montague RR (1997) A neural substrate of prediction and reward. *Science* 275: 1593–1599.
- Niv Y (2009) Reinforcement learning in the brain. *Journal of Mathematical Psychology* 53: 139–154.
- Wittgenstein L (1953) *Philosophical investigations*. Oxford: Blackwell. 250 p.
- Rosenblatt F (1959) Two theorems of statistical separability in the Perceptron. In: Blake V, Utley AM, eds. *Mechanisation of thought processes: Proceedings of a symposium held at the National Physical Laboratory*. London: Her Majesty’s Stationery Office. pp 419–456.
- Gallistel CR (2003) Conditioning from an information processing perspective. *Behavioural Processes* 61: 1234.
- Darwin C (1877) *Biographische Skizze eines kleinen Kindes*. *Kosmos* 1: 367–376.

35. Rice N (1980) *Cognition to language*. Baltimore: University Park Press. 203 p.
36. Landau B, Gleitman LR (1985) *Language and experience: Evidence from the blind child*. Cambridge: Harvard University Press. 256 p.
37. Roberson D, Davidoff J, Davies IRL, Shapiro LR (2004) The development of color categories in two languages: A longitudinal study. *Journal of Experimental Psychology: General* 133: 554–571.
38. Ramscar M, Dye M, Yarlett D (2009) No representation without taxation: The costs and benefits of learning to conceptualize the environment. *Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam, The Netherlands*.
39. Elman JL (1990) Finding structure in time. *Cognitive Science* 14: 179–211.
40. Ramscar M, Dye M, Gustafson J, Klein J (2009) Two routes to cognitive flexibility: Learning and response conflict resolution in the Dimensional Change Card Sort Task. *Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam, The Netherlands*.
41. Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* 268: 1632–1634.
42. Kamide Y, Altmann GTM, Haywood SL (2003) The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language* 49: 133–156.
43. Dahan D, Tanenhaus MK (2005) Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin and Review* 12: 453–459.
44. Fernald A, Perfors A, Marchman VA (2006) Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the second year. *Developmental Psychology* 42: 98–116.
45. Clark EV (2009) *First Language Acquisition*. Cambridge, UK: Cambridge University Press. 366 p.
46. Gentner D, Boroditsky L (2001) Individuation, relativity and early word learning. In: Bowerman M, Levinson S, eds. *Language acquisition and conceptual development*. Cambridge, UK: Cambridge University Press. pp 215–256.
47. Fernald A, Thorpe K, Marchman VA (2010) Blue car, red car: Developing efficiency in online interpretation of adjective-noun phrases. *Cognitive Psychology* 60: 190–217.
48. Thorpe K, Fernald A (2006) Knowing what a novel word is not: Two-year-olds “listen through” ambiguous adjectives in fluent speech. *Cognition* 100: 389–433.
49. Baldwin DA (1995) Understanding the link between joint attention and language. In: Moore C, Dunham PJ, eds. *Joint attention: Its origins and role in development*. Hillsdale NJ: Lawrence Erlbaum. pp 131–158.
50. Carpenter M, Nagell K, Tomasello M (1998) Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age. *Monographs of the Society for Research in Child Development* 63: 1–166.
51. Triesch J, Teuscher C, Deák GO, Carlson E (2006) Gaze following: why (not) learn it? *Developmental Science* 9: 125–147.
52. Rowe ML (2008) Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language* 35: 185–205.
53. Davies M (2009) The 385+ Million Word Corpus of Contemporary American English (1990–present). *International Journal of Corpus Linguistics* 14: 159–90.
54. Davies M (2010) Creating Useful Historical Corpora: A Comparison of CORDE, the Corpus del Español, and the Corpus do Português. In: Enrique-Arias A, ed. *Diacronía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus*. Frankfurt, Germany/Madrid, Spain: Vervuert/Iberoamericana. pp 137–166.
55. Benford F (1938) The law of anomalous numbers. *Proceedings of the American Philosophical Society* 78: 551–572.
56. Revkin SK, Piazza M, Izard V, Cohen L, Dehaene S (2008) Does subitizing reflect numerical estimation? *Psychological Science* 19: 607–614.
57. Cordes S, Gelman R, Gallistel CR (2001) Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin and Review* 8: 698–707.
58. Sonnenschein S, Whitehurst GJ (1982) The effects of redundant communications on the behavior of listeners: Does a picture need a thousand words? *Journal of Psycholinguistic Research* 11: 115–125.
59. Spivey MJ, Tyler MJ, Eberhard KM, Tanenhaus MK (2001) Linguistically Mediated Visual Search. *Psychological Science* 12: 282–286.
60. Booth JL, Sigler RS (2008) Numerical Magnitude Representations Influence Arithmetic Learning. *Child Development* 79: 1016–1031.
61. Jordan NC, Glutting J, Ramineni C (2010) The importance of number sense to mathematics achievement in first through third grades. *Learning and Individual Differences* 20: 82–88.
62. Jordan NC, Kaplan D, Ramineni C, Locuniak MN (2009) Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology* 45: 850–867.
63. Clements DH, Sarama J (2007) Effects of a Preschool Math Curriculum: Summative Research on Building Blocks Project. *Journal for Research in Mathematics Education* 38: 136–163.
64. Ramscar M, Suh E, Dye M (2011) How pitch category learning comes at a cost to absolute frequency representations. *Proceedings of the 33rd Meeting of the Cognitive Science Society, Boston, MA*.
65. Rescorla RA (1968) Probability of Shock in the Presence and Absence of CS in Fear Conditioning. *Journal of Comparative and Physiological Psychology* 66: 1–5.
66. Roberts S, Pashler H (2000) How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107: 358–367.