Received: 4 June 2019 | Revised: 17 December 2019 | Accepted: 16 January 2020

DOI: 10.1111/desc.12940

PAPER

Developmental Science WILEY

Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics

Alberto Testolin[1,2] | Will Y. Zou[3] | James L. McClelland[4]

[1]Department of General Psychology, University of Padova, Padova, Italy

[2]Department of Information Engineering, University of Padova, Padova, Italy

[3]Department of Electrical Engineering, Stanford University, Stanford, CA, USA

[4]Department of Psychology, Stanford University, Stanford, CA, USA

Correspondence

Alberto Testolin, Department of General Psychology, University of Padova, Via Venezia 12, Padova 35131, Italy.
Email: alberto.testolin@unipd.it

James L. McClelland, Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA.
Email: jlmcc@stanford.edu

Funding information

Fondazione Cassa di Risparmio di Padova e Rovigo; Università degli Studi di Padova

Abstract

Both humans and non-human animals exhibit sensitivity to the approximate number of items in a visual array, as indexed by their performance in numerosity discrimination tasks, and even neonates can detect changes in numerosity. These findings are often interpreted as evidence for an innate 'number sense'. However, recent simulation work has challenged this view by showing that human-like sensitivity to numerosity can emerge in deep neural networks that build an internal model of the sensory data. This emergentist perspective posits a central role for experience in shaping our number sense and might explain why numerical acuity progressively increases over the course of development. Here we substantiate this hypothesis by introducing a progressive unsupervised deep learning algorithm, which allows us to model the development of numerical acuity through experience. We also investigate how the statistical distribution of numerical and non-numerical features in natural environments affects the emergence of numerosity representations in the computational model. Our simulations show that deep networks can exhibit numerosity sensitivity prior to any training, as well as a progressive developmental refinement that is modulated by the statistical structure of the learning environment. To validate our simulations, we offer a refinement to the quantitative characterization of the developmental patterns observed in human children. Overall, our findings suggest that it may not be necessary to assume that animals are endowed with a dedicated system for processing numerosity, since domain-general learning mechanisms can capture key characteristics others have attributed to an evolutionarily specialized number system.

KEYWORDS

approximate number system, computational modeling, deep neural networks, number sense development, numerosity perception, visual number sense

1 | INTRODUCTION

It is well accepted that both humans and non-human animals are able to make approximate judgments of relative numerosity (Dehaene, 2011). Discriminability of two visual numerosities can be characterized, at least approximately, as a function of their ratio, in accordance with Weber's law (Dehaene, 2003). Notably, ratio-dependent performance has been observed also in infants (Xu, Spelke, & Goddard, 2005) and even neonates (Izard, Sann, Spelke, & Streri, 2009), although discriminability

Alberto Testolin and Will Y. Zou contributed equally to this work.

consistently improves with development over the period from infancy to adulthood (Halberda & Feigenson, 2008; Odic, Libertus, Feigenson, & Halberda, 2013; Piazza et al., 2010). For example, while 6-month-old infants can discriminate visual displays of dots with a ratio of 2:1 or greater, they fail to discriminate smaller ratios such as 3:2 (Xu et al., 2005). Discrimination performance is commonly quantified by estimating the subject's Weber fraction, a psychophysical measure which is thought to reflect the precision of the underlying numerosity representations (Halberda, 2011).

How should we understand these findings? One widely held view points to an innate, phylogenetically primitive system that has been called the 'Approximate Number System' (Feigenson, Dehaene, & Spelke, 2004). According to this view, the ability to estimate number enhances individual fitness (e.g. for finding abundant sources of food), and natural selection led to its early emergence and preservation across much if not all of the animal kingdom (Agrillo, 2014; Butterworth, 1999; Cantlon & Brannon, 2007; Ferrigno & Cantlon, 2017; Leslie, Gelman, & Gallistel, 2008; Nieder, 2005). However, the ability of organisms to respond to a particular dimension of variation in the environment might not be evolutionary pre-specified as such – instead, general purpose adaptive mechanisms might be sufficient, and these might easily be available to a wide range of species. In this article, we thus argue for an *emergentist view*, which emphasizes the possibility that cognitive abilities – including the capacity to make approximate judgements of numerosity – arise from the interplay between experience and domain-general learning mechanisms (McClelland et al., 2010). Indeed, the significant change in numerosity sensitivity observed during development suggests that experience and learning play key roles in shaping our numerosity representations.

Our emergentist framework can be naturally instantiated in the form of computational simulations based on artificial neural networks (Rumelhart & McClelland, 1986), which can reproduce elementary numerical abilities through hardwired mechanisms (Dehaene & Changeux, 1993) or, most interestingly, as a result of statistical learning over inputs of varying numerosity (Verguts & Fias, 2004). A crucial step forward in modelling number sense has been provided by more recent computational work based on deep neural networks (Cappelletti, Didino, Stoianov, & Zorzi, 2014; Chen, Zhou, Fang, & McClelland, 2018; Stoianov & Zorzi, 2012; Zorzi & Testolin, 2018). In particular, the seminal model of Stoianov and Zorzi (2012) simulated human-like numerosity judgments over two-dimensional displays that spanned a wide range of numerosities and incorporated variations in non-numerical visual properties. The model implemented a form of unsupervised representation learning (Bengio, Courville, & Vincent, 2012), where the objective is to create an invertible code on the internal (hidden) layers that can be used to accurately reconstruct the observed sensory input.

Here we further extend this computational approach by addressing three outstanding questions:

Research Highlights

- Although even newborns are sensitive to the approximate number of items in a visual display, our numerical acuity gradually improves during development.
- Initial sensitivity to numerosity can be simulated even in randomly initialized neural networks, and unsupervised deep learning leads to a progressive refinement of numerical representations.
- Numerosity sensitivity emerges in controlled environments (uniform number frequency and orthogonal variation between number and area) and with experience statistics mirroring that of natural environments.
- Our work suggests that domain-general learning mechanisms are sufficient to capture the key characteristics of the 'number sense' and its development.

1.1 | What is the initial numerical competence of deep neural networks?

The investigation of a fully trained network (Stoianov & Zorzi, 2012) does not address the crucial finding that even infants (including neonates) can exhibit a degree of sensitivity to differences in numerosity (Izard et al., 2009; Xu et al., 2005). In our simulations we therefore explore how well the initial state of a deep network can support numerosity discrimination. Surprisingly, it turns out that even randomly connected deep networks can support numerosity judgments, thereby shedding light on what might be a sufficient neural architecture not only to account for adult competence, but also to address the remarkable sensitivity to numerosity exhibited by human newborns.

1.2 | What is the developmental trajectory of number sense in deep neural networks?

Unsupervised deep learning models are usually trained in a 'greedy layer-wise' fashion (Hinton & Salakhutdinov, 2006), where learning is completed at one layer of the hierarchy (starting with the one closest to the input) before progressing to the next deeper layer. In line with previous developmental modelling (McClelland, 1989, 1994; Rogers & McClelland, 2004; Seidenberg & McClelland, 1989), we instead explore the possibility that acquisition of numerical acuity might arise from a gradual learning process. To this aim, we formulate a novel progressive algorithm for unsupervised deep learning, which allows adjustment of all of the connection weights following each sensory experience. Our developmental simulations show that this learning regimen supports a gradual improvement of numerical acuity, in line with experimental findings on humans. Moreover we provide a better quantitative characterization of the developmental trajectories of numerical acuity, incorporating the initial numerical competence measured at birth (for children) or at initialization (for our deep networks).

1.3 | Would a number sense emerge in deep neural networks exposed to the statistical structure of natural environments?

A key principle in computational neuroscience is that perceptual systems are adapted to the statistical properties of the surrounding environment (Fiser, Berkes, Orbán, & Lengyel, 2010; Girshick, Landy, & Simoncelli, 2011). This is also a foundational principle of unsupervised deep learning, where the objective is to discover high-order statistical structure that captures the distribution of the training data (Hinton, 2007). The model of Stoianov and Zorzi (2012) was trained with synthetic images where number and cumulative area were orthogonally varied, and where all numerosities appeared with the same frequency. Such a training corpus does not reflect the statistical structure of natural environments, where cumulative area might co-vary with number and where the frequency distribution of numerosities is far from uniform. We thus investigate how the statistics of natural environments might affect the emergence of number sense in deep networks by training the model using numerosity, size and position information from a large-scale natural image corpus. Although the use of different corpus statistics indeed modulates the learning outcome, our results show that the computational architecture itself may play an important role in the emergence of a Weber-like encoding of numerosity information, offering a distinct and viable alternative to other approaches for understanding the compressed encoding of numerosity in real brains (Piantadosi, 2016).

2 | METHODS

In this section, we will first introduce the motivation and the general structure of our progressive deep learning algorithm. We will then describe the materials and procedures used for training and testing the models. Additional technical details about network architecture, learning hyperparameters and model fitting are provided in the Supporting Information.[1]

2.1 | A developmental approach for unsupervised deep learning

The neural network model we use grows out of the proposal that perceptual processing takes place in a hierarchical processing system, with neurons in the lower layers encoding simple visual properties that are successively combined into more complex features (Fukushima, 1980; McClelland & Rumelhart, 1981; Riesenhuber & Poggio, 1999). Subsequently, it was demonstrated that 'deep' neural networks composed of several layers of non-linear processing units can be effectively trained with the simple objective of reconstructing their own input (Hinton & Salakhutdinov, 2006). This approach is often called *unsupervised deep learning* since the deep network itself is not explicitly trained to represent numerosity or other specified characteristics of the input (Hinton, 2007; Testolin & Zorzi, 2016) – the only constraint is to form an internal representation that minimizes error in reconstructing the inputs included in the training data. We view these models as widely applicable to modelling perception and perceptual learning, since the bi-directional propagation of activity in these models allows them to combine bottom-up sensory information with top–down expectations (McClelland, 2013) and shows how abstract representations can emerge through unsupervised learning (Zorzi, Testolin, & Stoianov, 2013).

Our model is based on a stack of two auto-encoders (see Supporting Information). In order to explore the developmental time-course of deep learning, we develop a progressive alternative to the greedy layer-wise training approach commonly used in machine learning (ML) as discussed above. The work of Stoianov and Zorzi (2012) employed the greedy layer-wise approach, in keeping with their goal of modelling the adult performance. In our developmental approach, instead, we interleave learning in each layer-wise stage so that each is performed in the course of processing each training example (Figure 1). Besides extending the use of neural networks previously used to capture the acquisition of semantic, perceptual and other cognitive abilities (McClelland, 1994; Seidenberg & McClelland, 1989), our scheme is consistent with the complementary learning systems theory (Kumaran, Hassabis, & McClelland, 2016; McClelland, McNaughton, & O'Reilly, 1995), where perceptual and cognitive abilities gradually arise through the accumulated impact of adjustments to connections made after each learning experience.

As illustrated in Figure 1, for each training example presented to the network, our algorithm combines one iteration of layer-wise training with one iteration of full network reconstruction. Note that learning remains completely unsupervised, in the sense that the learning task is only to build an internal model that optimizes the network's ability to reconstruct the input patterns. We followed Stoianov and Zorzi (2012) in using two hidden layers with 80 units in the first layer and 400 in the second. For each learning trial, a random item was drawn with replacement from the training data set, and the developmental algorithm described above was used to update the weights, using a constant learning rate of 0.01, with no weight decay or momentum. Training continued for 2,800,000 pattern presentations, a regime that would correspond to approximately 300 patterns per day over a 25-year period.

2.2 | Visual stimuli

Stimuli were presented to the network as two-dimensional images of size 30 × 30 pixels containing a different number of white rectangles drawn on a black background. We trained the deep networks on

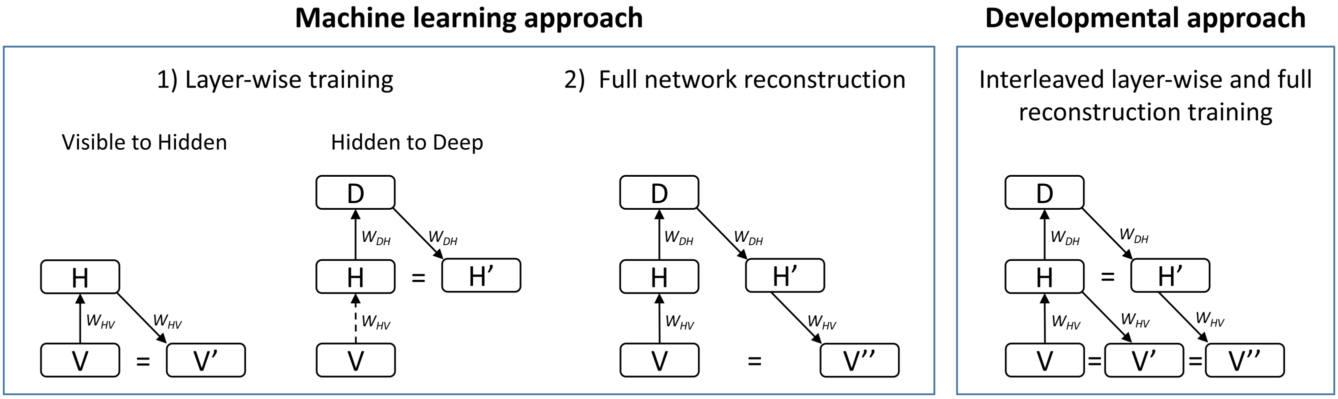
Machine learning approach
Developmental approach
1) Layer-wise training
2) Full network reconstruction
Interleaved layer-wise and full reconstruction training
Visible to Hidden
Hidden to Deep
D
H
H'
V
V'
V''
=
W_{DH}
W_{HV}

FIGURE 1 Comparison of the standard machine learning (ML) approach with our developmental approach to training a deep neural network. The objective was to create an invertible code D at the deepest layer, optimized to allow an input V to be reproduced at the network's output (V′ or V′′).In the ML approach, the connection weights between the hidden and visible layers WHV are learned in a first stage of training over the full set of training examples; weights are optimized to learn the invertible mapping that makes V′ as similar to V as possible, as indicated by the "=" sign. These weights are then frozen (dotted arrow in the mid panel), and the weights between the deep and hidden layer WDH are optimized to make H′ as similar to H as possible. In a final fine-tuning stage, the weights are further optimized to make V′′ as similar as possible to V after propagating activation from the input to the deep layer and back. In our developmental approach, one iteration of each of the three steps is instead conducted in the course of processing each pattern

three data sets (samples of stimuli are shown in Figure 2a, and their statistical properties are reported in Figure 3), in order to investigate how different characteristics of the training corpus could have an impact on the time course of learning.

The first one, which we call the 'S&Z data set', was generated using the procedure described in Stoianov and Zorzi (2012). Two factors (visual numerosity and cumulative surface area) were varied in order to create rectangular white patches to be placed in the display. In particular, for each pattern, the cumulative area was randomly sampled from a uniform discrete distribution in steps of 32 between 32 and 256 pixels, while the target numerosity was randomly sampled from a uniform distribution between 1 and 32.

The second data set, which we call the 'Natural data set', was generated from popular computer vision data sets used for the PASCAL detection challenge (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010), which consist of images with rectangular boxes indicating sizes and positions of objects. For each original image (see examples in Figure 2b), we replaced each object with its localized bounding box, projecting them as non-overlapping white rectangles in a black 30 × 30 pixel background while preserving the placement of objects to the extent possible. The number of objects per display ranged from 1 to 56, though the proportion of displays containing a given number of items fell off very quickly as the number of items increased.

Because the relationship between probability and n is often approximated by a power law of the form $p(n) \propto \frac{1}{n^{\alpha}}$, we show the resulting function of this form in Figure 3 (black dashed line; $\alpha = 2$). In our natural data set, the best-fitting value of α was 2.8. It is evident, however, that the fall off in $p(n)$ is more gradual than the formula implies for n <5, and steeper than this for n >10. The shape of the curve is thus better explained by a shifted power function $p(n) \propto \frac{1}{(\beta+n)^{\alpha}}$ (Mandelbrot, 1953; Piantadosi, 2014), also shown in Figure 3 (blue dashed line; $\alpha = 9.5$ and $\beta = 15$).

The third data set, which we call the 'Irregular-shape alpha = 2 (ISA2) data set', was constructed to address the extreme roll-off in relative frequency with n >10 in the natural data set, and to explore the possibility that, during development, the deep network might be exposed to visual stimuli containing irregular shapes rather than just rectangular and square items. To this aim, visual stimuli in the ISA2 data set contained ellipsoids of varying aspect ratios, whose per-item size distribution was approximately matched to that of the natural data set and whose number frequency distribution was defined according to a power law with $\alpha = 2$ (third column in Figure 3). This value seems to provide a good approximation to estimates of frequencies of occurrence of numbers in text (Dehaene & Mehler, 1992; Piantadosi, 2016).

2.3 | Testing procedure

Our interest in testing the model focused on assessing how well the internal representations on the deepest layer could support the kind of numerosity judgment assessed in behavioural studies. We should distinguish between the statistical properties of the materials used to *train* the deep network and the properties of those used during behavioural *testing*. We treat the deep network as a system that learns over developmental time from the statistics of its experiences. As such, we explore the effects of varying the statistical properties of these experiences on the numerosity sensitivity that the deep network can support, with the aim of simulating the influence of environment on development. When it comes to testing the network, we turn our attention instead to the characteristics of the materials used in behavioural assessments of human performance, with the aim of evaluating the model on a set of stimuli reflecting the statistical properties of those commonly employed in psychophysical experiments. Among

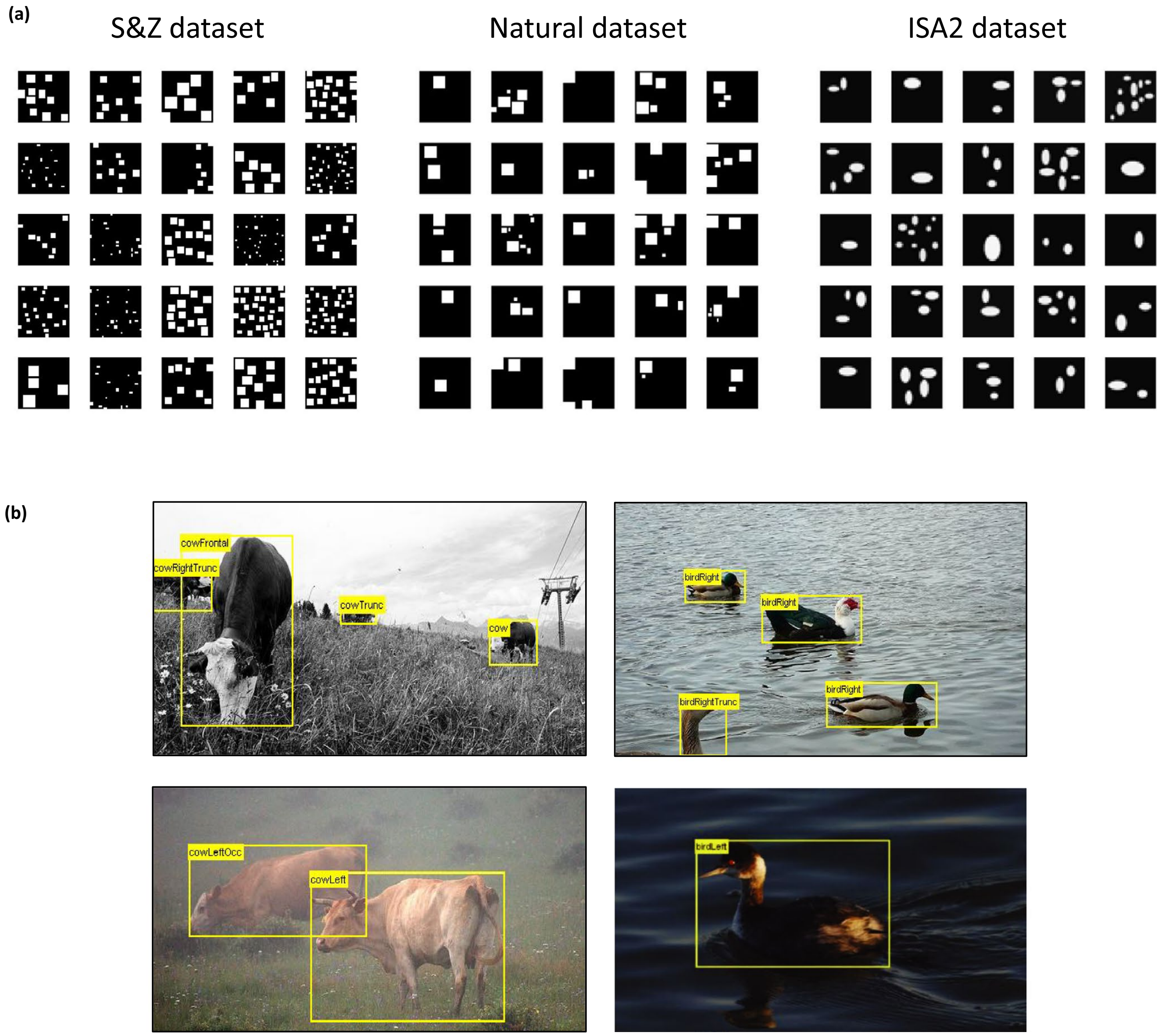

FIGURE 2 (a) Examples of visual stimuli from the three different data sets used in our simulations. (b) Examples of images from the PASCAL object detection challenge used to create the natural data set

other things, evidence suggests that in natural experience smaller numbers are encountered more frequently, but this is generally not true during behavioural testing, where experimenters use equal-frequency sampling of a range of numerosities. Therefore, although we trained the deep networks using each of the three data sets mentioned above, only the S&Z training and test data sets were used to train and test the classifier, since these stimuli have a balanced frequency distribution across numerosities, control for cumulative area, and use uniform random placement of items in the visual field.

In order to translate the network's internal representations into behavioural responses, we trained a simple classifier to map the internal representations at the deepest layer of the network into a binary classification response (Stoianov & Zorzi, 2012) indicating whether or not the visual input numerosity was larger than a reference number (two reference numbers were used, $N = 8$ and $N = 16$). We do not intend the classifier training as a model of the process whereby human subjects learn to map internal representations of numerosity onto overt responses. Instead, we treat the classifier as a procedure for allowing us to estimate the internal Weber fraction of the model, that is, to measure how well the deep networks' representations can support explicit numerosity judgments (for discussion, also see Zorzi et al., 2013). The classifier weights were optimized for each of the initial weight conditions and at each of several different time points during developmental training, in order to assess the representational capacity of the deep network under each initialization or developmental stage. We report results based on 10 simulation runs using different random initial weights and random sequences of training examples to train the deep networks.

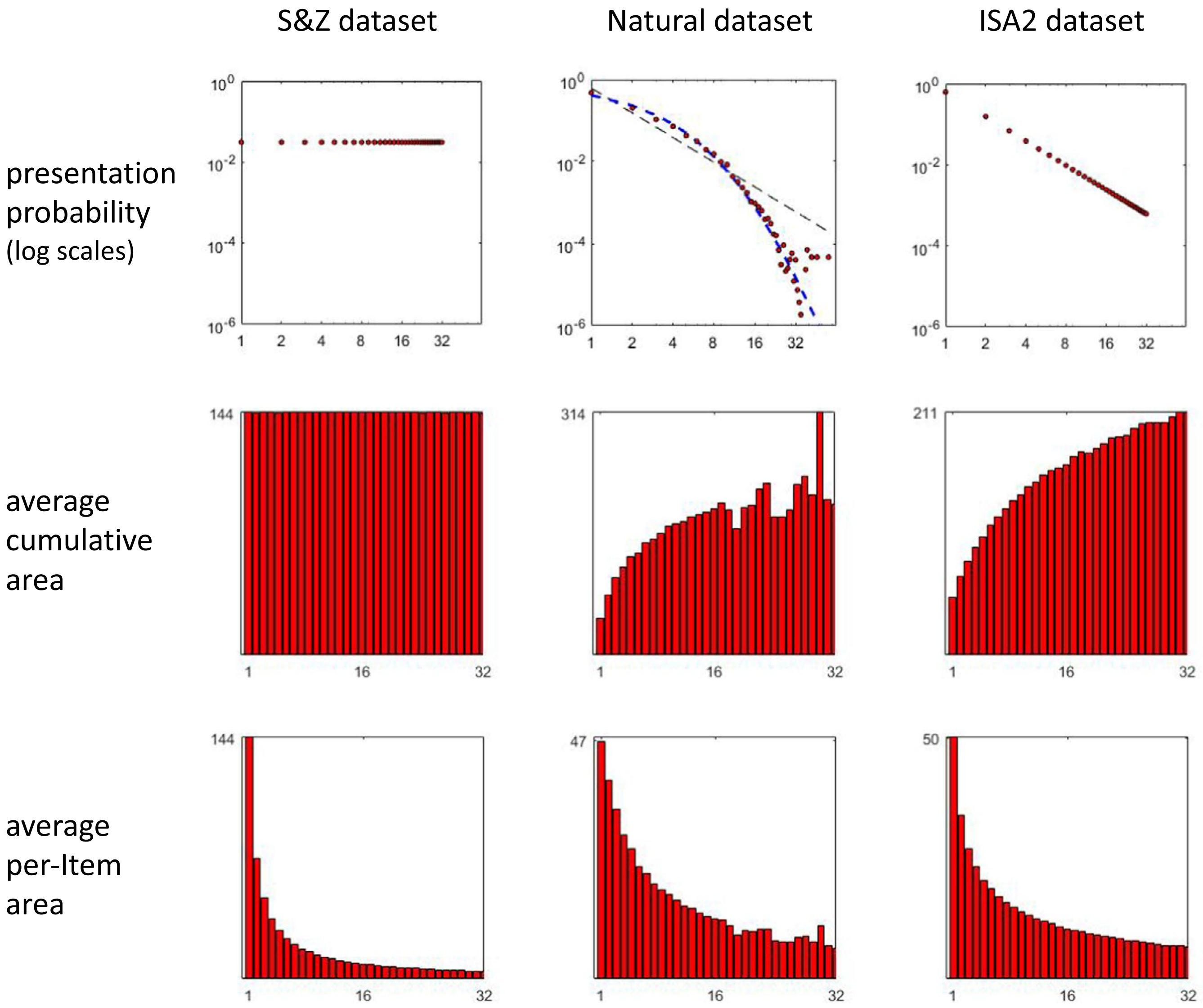

FIGURE 3 Statistics for the different data sets considered. For each numerosity, the graphs show the relative presentation probabilities plotted on a log-log scale (top panels), the average cumulative area of all the items in the displays (middle panels) and the average area of individual items in the displays (bottom panels). In the presentation probability graph for the natural data set, the dashed black line (Zipfian distribution with alpha = 2) highlights that the images in the natural data set exhibit a steeper decrease in number frequencies. The trend is better captured by the dashed blue line (shifted Zipfian distribution with alpha = 9.5 and beta = 15)

2.4 | Weber fraction estimation

The response distribution (probability of 'larger' responses) of the classifier resulted in a psychometric function (see Figure S3) which was used to estimate the model's Weber fraction. The primary data we compare our model to comes from experiments in which participants compared two numerosities and indicated which one was larger. One of these stimuli is usually based on a reference numerosity n_r (for example, Piazza and colleagues used 16 and 32 as reference numbers) and the other stimulus has a different, comparison numerosity n_c. The psychophysical model standardly thought to underlie such comparisons holds that the data is based on independent samples from two Gaussians, one for each of the two displays. Under the logarithmic Gaussian model of numerosity representation (Dehaene, 2012), each distribution is treated as having a mean equal to the log of the presented numerosity, and both are assumed to have the same standard deviation *w*. When asked to determine which stimulus has the larger numerosity, the participant is thought to compare a sampled value s_c from the distribution of mean n_c, to a sampled value s_r from the distribution of mean n_r, and to designate the comparison stimulus as the larger of the two if $s_c > s_r$. The probability that $s_c > s_r$ is equivalent to the probability that the difference $s_c - s_r$ is greater than 0, and this difference is also a normally distributed random variable with mean log(n_c) − log(n_r) and standard deviation $\sqrt{2}w$. Given this, the probability of choosing the comparison stimulus should be:

$$p\left(s_c > s_r\right) = 1 - \Phi\left(\log\left(n_c/n_r\right), \sqrt{2}w\right),$$

where Φ is the probability density function of the Normal distribution, evaluated at *x* = 0. This curve can then be fit to produce an estimate of the Weber fraction w, as was done in Piazza et al. (2010). However, the √2 scale factor is omitted in our case, since the classifier decision is taken by comparing a single sample from the distribution of the comparison numerosity with a fixed internal reference (either $N = 8$ or $N = 16$), so that all the variability would only be in the estimate of the numerosity of the display (see Supporting Information for more details). This choice is conservative, in the sense that it results in larger estimates of w than would be obtained if the √2 factor was included, as it was in the estimation of w by Stoianov and Zorzi (2012); thus, for the same observed pattern of classification accuracy (Figure S3), our estimated value of w will be larger than the value they reported.

3 | RESULTS

We first examined how well the classifier could perform based on representations formed in the deep network with untrained (random) initial weights. This allowed us to consider whether a generic neural network could support numerosity judgments even without any prior exposure to environmental stimulation. In addition, this allowed us to examine whether variations in the range of the initial weights might affect initial numerosity sensitivity, considering this to be the kind of parameter that might be optimized by evolution. Others have shown that networks with random weights can support surprisingly good performance in a variety of classification tasks if the range of weight values is optimized (Jaeger, Maass, & Principe, 2007; Widrow, Greenblatt, Kim, & Park, 2013). We therefore randomly initialized the weight matrices of the network using different values of the random weight standard deviation parameter σ, setting unit biases of all layers to zero. Using representations produced at the second layer of such random networks, the Weber fraction was computed as the average value resulting for each of 10 different randomly initialized networks for each value of σ.

Results in the first row of Table 1 show that if the standard deviation of the initial connection weights was set to 0.1 or 0.5, random initialization of the two-layer neural network supported a Weber fraction of about 1.6. Smaller or larger standard deviation values generally led to worse performance. Furthermore, the second row of the table illustrates an interesting finding from a variant of the same simulation: if only the first layer of the network is fully trained on the S&Z data set, a Weber fraction around 0.4 (rivaling that of children over 4 years old) can be obtained from the output of the second layer, over a wide range of values of the standard deviation used to initialize the second layer weights. These findings are aligned with computational studies showing that random matrices can support approximate encoding of magnitudes (Hannagan, Nieder, Viswanathan, & Dehaene, 2018), and with recent work carried out with more sophisticated architectures showing that numerosity sensitivity can be observed in untrained convolutional neural networks (Kim, Jang, Baek, Song, & Paik, 2019). One possible explanation for such findings is that random weights can provide an approximate signal that co-varies with continuous visual features (e.g. cumulative area, item size, density, etc.) that are correlated with number. Indeed, it has been shown that deep networks endowed with basic visuospatial processing can exhibit a remarkable accuracy in numerosity discrimination, but only when numerosity covaries with other magnitudes (Zorzi & Testolin, 2018), and in preliminary investigations of our networks we observed a similar effect. Notably, a similar influence of continuous visual features is also observed in numerosity judgments in younger human age groups (Halberda & Feigenson, 2008; Soltész, Szűcs, & Szűcs, 2010).

Based on the results obtained with random initializations, we set the connection weight initialization parameter σ to 0.1 for the developmental simulations. Learning trajectories of the deep networks trained on the three different data sets are reported in Figure 4, with error bars representing the standard deviation of the mean for the 10 different random initializations. For all data sets the developmental trajectory followed a similar pattern, exhibiting rapid learning at first followed by a levelling off towards the end of training. Learning progressed at a somewhat smaller rate in the deep networks trained on the ISA2 data set (green curve): this phenomenon is not likely to be due to the Zipfian distribution of numerosities in that corpus, since a similar distribution is also present in the Natural data set (blue curve). We hypothesize that this might instead be due to the less defined borders of the objects, which might have caused a slower development of localized feature detectors in the first hidden layer. However, the Zipfian distribution of numerosities does appear to have an impact on the endpoint of learning, since the final acuity of the deep networks trained on the ISA2 and Natural corpora is worse than the final acuity of the networks trained on the S&Z corpus (red curve).

Figure 5 shows the individual data points and the developmental trajectories obtained from three longitudinal behavioural studies (Halberda & Feigenson, 2008; Odic et al., 2013; Piazza et al., 2010). We also included two additional data points derived from experiments

TABLE 1 Weber fractions for untrained deep networks by level of variability in initial weights

Architecture	σ = 10	σ = 1	σ = 0.5	σ = 0.1	σ = 0.01
Both layers randomly initialized	1.94 ± 0.23	1.65 ± 0.13	1.60 ± 0.13	1.58 ± 0.17	4.55 ± 1.10
Only second layer randomly initialized	0.37 ± 0.01	0.38 ± 0.01	0.37 ± 0.01	0.37 ± 0.01	0.38 ± 0.01

Note: The parameter σ represents the standard deviation of the initial random weights. The numbers in each cell are the mean of the value of the Weber fraction fitted to the classifier output ± the standard deviation of the fitted values, based on 10 separate networks initialized using the indicated value of σ.

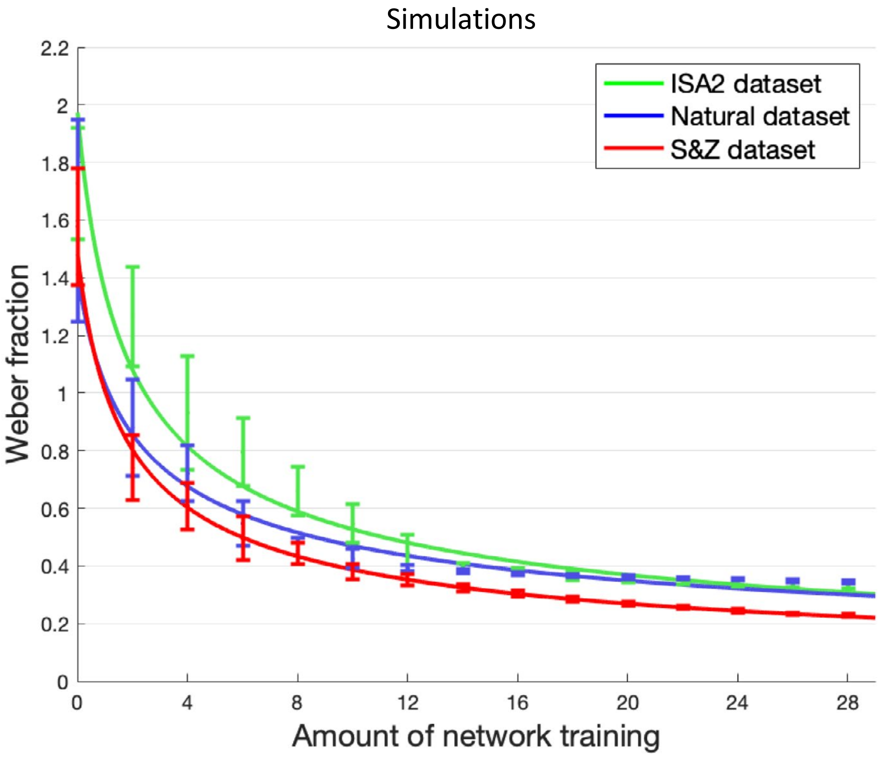

FIGURE 4 Developmental trajectories for deep networks trained using the three different data sets. The first plotted data point reflects performance prior to any learning (i.e. random initialization), while successive points thereafter are at intervals of 200,000 pattern presentations. Fitted curves correspond to the power model with initial competence

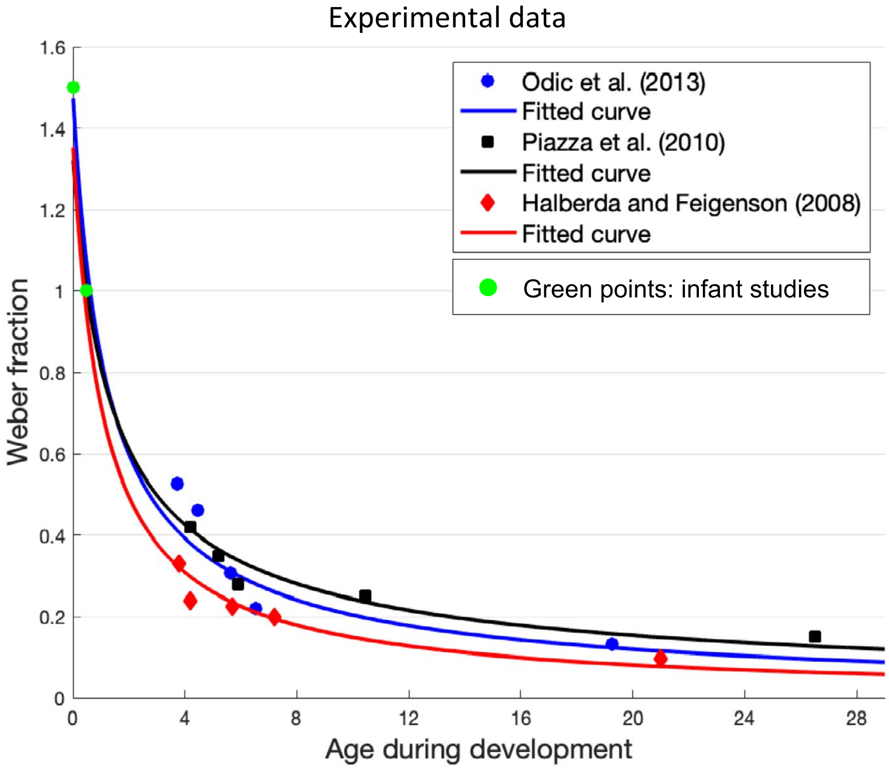

FIGURE 5 Developmental trajectories derived from experimental data reported in three independent studies. Fitted curves correspond to the power model with initial competence

with infants, which used coarse estimates (see Supporting Information) of the Weber fraction right after birth (Izard et al., 2009) and at the age of 6 months (Xu et al., 2005). Although these estimates are based on habituation rather than a two alternative forced choice, they might still provide a useful indication of initial numerosity sensitivity, and such estimates have also been included in previous attempts to characterize the human developmental trajectory (Halberda & Feigenson, 2008; Odic et al., 2013).

Overall, the simulations start at an average Weber fraction of 1.6, a value compatible with the value of 1.5 we estimated for newborns.

The final Weber fractions achieved by the deep networks were between 0.2 and 0.4. These values fall within the range of the mean values across participants estimated by other studies on human adults (DeWind, Adams, Platt, & Brannon, 2015; Halberda, Ly, Wilmer, Naiman, & Germine, 2012; Halberda, Mazzocco, & Feigenson, 2008; Piazza, Pica, Izard, Spelke, & Dehaene, 2013; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008), although values as high as 0.4 are at the high end of the adult range. For reasons that are not fully clear, estimated values of *w* for adults reported from developmental studies tend to fall below the final values measured in the studies mentioned above; for example, a value of 0.15 was reported for the educated Italian adult group in the developmental study of Piazza et al. (2010).

To provide a descriptive quantitative characterization of both the simulated and empirical developmental curves, we fitted three different descriptive functions to each series of data points. The simplest function was simply a power law function in the form:

$$w = ax^{-b},$$

where *w* is the predicted value of the Weber fraction, *x* is the age during development, and *a* and *b* are two parameters that represent respectively the normalization constant and the exponent of the function. This is the function that has been traditionally used to characterize the developmental trend of numerical acuity (Halberda & Feigenson, 2008; Odic et al., 2013), and we call it the 'simple power model'. One limitation of this model is that it assigns extremely large values (approaching infinite) to the Weber fraction within hours and days of birth. To address this, we introduce an initial constant to produce a value for the Weber fraction 'at birth', producing the equation:

$$w = a\left(1 + x\right)^{-b}.$$

We call this the 'power model with initial competence'. This equation imposes an arbitrary scaling of the relative importance of initial competence compared to subsequent development. To allow for flexible weighting of these two factors, we also introduce an additional parameter *s* that serves as a scaling factor. The equation therefore becomes:

$$w = a\left(1 + sx\right)^{-b}.$$

We call this the 'power model with initial competence, scaled'. This additional parameter increases the flexibility of the model, at the cost of one fewer degree of freedom in the fit.

Using a maximum likelihood fitting procedure, the best fitting models were those taking into account initial competence, both for the empirical and for the simulations (see Supporting Information for details). The fitted curves shown in Figures 4 and 5 are based on the power model with initial competence, which provided a very accurate fit (see Figure S4 for plots of all other models, and Table S1 for the corresponding estimated parameters[2]). Overall, the shape of the simulated learning curves is strikingly similar to that of the empirical developmental trajectories, as also attested by the parameters resulting from the fitting procedure (see Table 2), though the exponent estimated for the simulated curves is generally smaller.

4 | DISCUSSION

Overall, our computational simulations show that (a) the initial numerosity discrimination ability of randomly initialized deep networks could rival that of human newborns; (b) their gradual development follows trajectories very similar to those observed in human longitudinal studies; and (c) their final competence approximates that of human adults. These results are consistent with an emergentist perspective in which generic properties and learning mechanisms of neural systems might be sufficient to characterize the ability to approximately represent and process numerosity information.

Regarding the first point, it should be noted that connection weights need to be initialized within a well-chosen range of values. While we made this choice to optimize numerosity performance, a

TABLE 2 Results of model fitting for empirical and simulated developmental trajectories using maximum likelihood estimation, for the 'power law with initial competence'

	Power model with initial competence					
	a	*b*	r^2	*NLL*	χ^2	*df*
Empirical data						
Odic et al. (2013)	1.49	0.83	0.98	0.17	7.46	5
Piazza et al. (2010)	1.33	0.71	0.98	0.09	3.43	5
Halberda and Feigenson (2008)	1.37	0.92	0.98	0.11	4.54	5
Simulation data						
S&Z data set	1.49	0.56	0.99	0.02	0.22	13
Natural data set	1.42	0.46	0.97	0.11	4.58	13
ISA2 data set	1.98	0.55	0.94	0.14	4.88	13

Note: Tables report the estimated values for the functional parameters (*a* and *b*), the resulting statistics (r^2, Negative Log Likelihood, χ^2) and the corresponding degrees of freedom (*df*).

similar choice might be made by evolution for other reasons. Indeed, considerations applicable to generic multi-layer networks (Glorot & Bengio, 2010) suggest that weights initialized between our two best values (σ = 0.1 and 0.5) would be optimal for our networks.[3] Thus, only very generic assumptions are required to create a network supporting a discrimination ability at a level close to that of infants.

Regarding the second point, our deep networks developed representations that gradually became more and more refined, supporting greater and greater acuity in approximate numerosity judgments following a progressive refinement similar to that seen over human development. Initial discriminability improves rapidly and then levels off, with much slower progress in later stages of learning. Such a pattern is widely observed in developmental and learning studies across a wide range of domains and it is captured, for reaction time data, in the famous *power law of practice* widely applied to behavioural data sets starting almost a century ago (Snoddy, 1926). The power law of practice is also observed in neural network models trained with back propagation (Cohen, Dunbar, & McClelland, 1990) as well as other models simulating learning dynamics (Newell & Rosenbloom, 1981). Given the ubiquity of power laws in both behaviour and simulation models of learning, we argue that this general pattern can be seen as an example of a generic characteristic of learning as a function of experience, rather than a signature of a special characteristic of a specific system for the representation of approximate numerosity.

Our research introduces a further refinement to the power law formulation used in numerosity perception research, allowing for initial competence prior to any actual experience, as exhibited by our randomly initialized networks. A similar allowance for initial competence is generally included in power law models of practice in other domains (e.g. Newell & Rosenbaum, 1981), and is necessary to describe the human data if initial competence at birth as observed by Izard et al. (2009) is to be captured. The initial competence can be viewed as representing time before birth during which the initial Weber fraction decreases from discriminability of 0 (corresponding to w = infinity) to its value at birth, idealizing the processes that take place *in utero* as neural networks are formed during embryological and foetal development (notably, it is known that random neural activity can contribute to the structuring of such networks even prior to birth, see, for example, Miller, Keller, & Stryker, 1989).

One of the purposes of our simulations was also to investigate how numerical acuity develops under different training distributions, exploring the question of whether a similar progressive trend would be observed regardless of the specific statistical properties encoded in the learning environment. Interestingly, the answer appears to be *yes*, since all models exhibited similar developmental trajectories. However, it should be stressed that the dataset derived from the statistics of photographs of natural scenes is only a tentative approximation of children's perceptual environment, which is also determined by the concurrent development of attention mechanisms allowing selection of the most relevant information. A promising venue for future research would thus be to investigate more fully how children parse egocentric visual scenes, in order to understand which (and how many) objects might be actually perceived at different developmental stages (Clerkin, Hart, Rehg, Yu, & Smith, 2017).

Regarding the final performance, all models achieved a numerical acuity well aligned with that reported in studies on human adults, though we should be cautious in performing quantitative comparisons. Indeed, it is well-known that different experimental protocols and testing conditions (e.g. lab settings vs. on-line crowdsourcing platforms) induce differences in the estimated Weber fraction, leading some authors to question the inter-test reliability of this measure (Clayton, Gilmore, & Inglis, 2015; Guillaume & Van Rinsveld, 2018; Price, Palmer, Battista, & Ansari, 2012). Furthermore, variability in individual estimates is very high even within the same experimental study, for example ranging from 0.15 to 0.3 in Revkin et al. (2008), from 0.1 to 0.5 in Halberda et al. (2008), from 0.1 to 0.4 in Piazza et al. (2013) and from 0.15 to 0.5 in DeWind et al. (2015). It should also be noted that our modeling captures only the process whereby experience leads to improvements in numerosity sensitivity. Factors other than those at work in our present model would have to be introduced to explain the reduction in numerosity sensitivity observed in older participants (Cappelletti et al., 2014; Halberda et al., 2012).

Interestingly, the deep networks trained on the S&Z data set achieved a slightly superior numerical acuity, possibly approaching that of educated humans, while the numerical acuity of the networks trained on the more ecological data sets could be more aligned towards that of uneducated populations (Piazza et al., 2013). On the one hand, the fact that adherence to Weber's law is observed even when the network is trained using a flat number frequency distribution (S&Z data set) suggests that adherence to Weber's law might not need to be explained at a 'rational analysis level', under the assumption that the brain maximizes average precision in numerosity representation when number frequency of occurrence decreases according to a power law (see Piantadosi, 2016, and other papers cited therein). On the other hand, the lower numerical acuity achieved by the networks trained on the data sets featuring a more naturalistic statistical structure is aligned with the finding that culture and formal education may play an important role in sharpening numerosity representations (Piazza et al., 2013). In this respect, the more controlled statistical structure of the S&Z data set might have encouraged the network to encode numerosity while avoiding confounding factors that orthogonally varied with number (such as cumulative area), a constraint that might somehow emulate the more extensive experience with number occurring in formal educational settings.

It is also important to emphasize that the classifier used to assess the representational competency of the deep network *is* trained to make numerosity judgments. In this respect, our procedure for training the classifier can be seen as aligned with the explicit feedback that is commonly used in empirical studies (Halberda & Feigenson, 2008; Izard & Dehaene, 2008; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Revkin et al., 2008). While some authors argue that the role of feedback is mostly motivational, as it helps in maintaining a high level of attention (Cantlon et al., 2009; Odic et al., 2013), others have shown that it also significantly improves numerosity sensitivity (DeWind & Brannon, 2012). It could also be argued that demonstrations of numerosity sensitivity right after birth (Izard et al., 2009) in a situation where no overt response is required, still provide a basis for thinking there may be initial representation of numerosity per se, since it is difficult to argue that children this young could have received classifier training. We acknowledge this possibility, although we note that recent simulations have shown that good discrimination performance can be achieved even when the classifier is trained with minimum supervision (Zorzi & Testolin, 2018).

We finally note that children's numerosity judgments can be strongly influenced by item size (Soltész et al., 2010) or other non-numerical visual properties (Clayton et al., 2015), and studies that find lesser influence of such factors provide children with trial-by-trial accuracy feedback (Halberda & Feigenson, 2008; Odic et al., 2013), essentially allowing the experiment itself to provide a classifier training signal. Interestingly, deep networks also exhibit sensitivity to congruency manipulations, for example by more frequently misclassifying stimuli where continuous visual cues are in disagreement with numerosity (Testolin, Dolfi, Rochus, & Zorzi, 2019; Zorzi & Testolin, 2018). While further work is needed to better establish how a generic neural network might support numerosity sensitivity in infancy, it could still be argued that explicit numerosity judgments may require feedback-driven tuning of classification responses in humans, as in our models.

A further possibility worth considering is the idea that experience with number itself might shape, not only the readout of internal representations, but also the representations themselves. In the current work, we have not relied on this possibility, but other neural network modeling work has shown that smaller values of *w* can be achieved in a network trained end-to-end with backpropagation to estimate the number of items in a display (Chen et al., 2018). Because numerical acuity improves over development even in uneducated individuals from cultures without number words, we think it is likely that the emergence of representations capable of supporting numerosity judgements can and do arise through unsupervised learning. At the same time, it seems reasonable to be open to the possibility that supervised learning could play a role in further shaping numerical acuity in educated populations, and even the possibility that non-visual information about number could act as a supervisory signal to promote refinement of visual numerosity sensitivity in uneducated populations and animals.

Overall, we believe that the computational work presented here constitutes an important step toward a better understanding of the mechanisms underlying the progressive development of our visual number sense. Our primary modelling effort concerned the identification of sufficient conditions that would allow a neural network to support sensitivity to numerosity prior to any visual experience as well as a gradual development of numerical acuity resembling that observed in children. One exciting venue for future work will be to investigate whether the inclusion of additional architectural and learning constraints in the model could help in capturing the high variability observed in empirical studies, thus possibly enlightening the computational bases of individual differences in numerical acuity. However, we believe that the most challenging direction for future research will be to push our framework into uncharted territory, in order to establish whether our perspective on numerical development could also account for the emergence of more complex, high-level numerical skills, such as those required to represent and manipulate symbolic numbers (Leibovich & Ansari, 2016).

ACKNOWLEDGMENTS

We thank the members of the PDP Lab at Stanford for comments and suggestions as the work progressed. A.T. was supported by the STARS Grant 2018 'Deepmath' from the University of Padova and by the Cariparo Excellence Grant 2017 'Numsense'. A.T. also acknowledges the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data and source code that support the findings of this study are publicly available at the Open Science Framework (OSF): https://osf.io/h5pfm.

ORCID

Alberto Testolin https://orcid.org/0000-0001-7062-4861

ENDNOTES

[1] The complete source code of our developmental algorithm, along with the code used to generate the visual stimuli and test the deep networks, are made publicly available at the Open Science Framework: https://osf.io/h5pfm.

[2] Slightly better fit statistics were obtained when including the scaling factor, but this could reflect overfitting due to the increase in the number of free parameters. We chose to present the simpler model here because the additional scaling parameter can trade off with the other parameters with little change in the overall goodness of fit, rendering the best-fitting values of the parameters less informative.

[3] Given *ns* sending units and *nr* receiving units, an optimal starting point for learning is provided by values distributed uniformly in the range [−*r*, *r*] were $r=4*\sqrt{(6/(ns+nr))}$ (the scale factor of 4 adjusts for our use of sigmoid units rather than *tanh* units as described in the original analysis in Glorot and Bengio, 2010). The formula gives values of *r* = .31 for the weights between the input (*n* = 900) and first hidden layer (*n* = 80) and 0.45 for the weights between the first hidden and second hidden (*n* = 400) layers. The corresponding standard deviations for these two cases are 0.18 and 0.25 respectively, between the values of 0.1 and 0.5 used in our simulations.

REFERENCES

Agrillo, C. (2014). Numerical and arithmetic abilities in non-primate species. In R. Cohen Kadosh & A. Dawker (Eds.), *The Oxford handbook of numerical cognition* (pp. 214–236). New York, NY: Oxford University Press.

Bengio, Y., Courville, A., & Vincent, P. (2012). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1–34.

Butterworth, B. (1999). *The mathematical brain*. London, UK: Macmillan.

Cantlon, J. F., & Brannon, E. M. (2007). How much does number matter to a monkey (Macaca mulatta)? *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(1), 32–41. https://doi.org/10.1037/0097-7403.33.1.32

Cantlon, J. F., Libertus, M. E., Pinel, P., Dehaene, S., Brannon, E. M., & Pelphrey, K. A. (2009). The neural development of an abstract concept of number. *Journal of Cognitive Neuroscience*, *21*(11), 2217–2229. https://doi.org/10.1162/jocn.2008.21159

Cappelletti, M., Didino, D., Stoianov, I., & Zorzi, M. (2014). Number skills are maintained in healthy ageing. *Cognitive Psychology*, *69*, 25–45. https://doi.org/10.1016/j.cogpsych.2013.11.004

Chen, S. Y., Zhou, Z., Fang, M., & McClelland, J. L. (2018). Can generic neural networks estimate numerosity like humans? In T. T. Rogers, M. Rau, X. Zhu & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the Cognitive Science Society* (pp. 202–207). Austin, TX: Cognitive Science Society.

Clayton, S., Gilmore, C., & Inglis, M. (2015). Dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement. *Acta Psychologica*, *161*, 177–184. https://doi.org/10.1016/j.actpsy.2015.09.007

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160055. https://doi.org/10.1098/rstb.2016.0055

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(3), 332–410. https://doi.org/10.1037/0033-295X.97.3.332

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4), 145–147. https://doi.org/10.1016/S1364-6613(03)00055-X

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. New York, NY: Oxford University Press.

Dehaene, S. (2012). Symbols and quantities in parietal cortex: elements of a mathematical theory of number representation and manipulation. In P. Haggard (Ed.), *Sensorimotor foundations of higher cognition* (pp. 526–574). Oxford, UK: Oxford Scholarship Online.

Dehaene, S., & Changeux, J.-P.-J. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience*, *5*(4), 390–407. https://doi.org/10.1162/jocn.1993.5.4.390

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

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265. https://doi.org/10.1016/j.cognition.2015.05.016

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: Effects of feedback and training. *Frontiers in Human Neuroscience*, *6*, 1–10. https://doi.org/10.3389/fnhum.2012.00068

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes challenge 2010 results. *International Journal of Computer Vision*, *88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4

Feigenson, L., Dehaene, S., & Spelke, E. S. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Ferrigno, S., & Cantlon, J. F. (2017). Evolutionary constraints on the emergence of human mathematical concepts. *Evolution of Nervous Systems*, *3*, 511–521. https://doi.org/10.1016/b978-0-12-804042-3.00099-3

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130. https://doi.org/10.1016/j.tics.2010.01.003

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *202*(4), 193–202. https://doi.org/10.1007/BF00344251

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *9*, 249–256. https://doi.org/10.1.1.207.2059

Guillaume, M., & Van Rinsveld, A. (2018). Comparing numerical comparison tasks: A meta-analysis of the variability of the weber fraction relative to the generation algorithm. *Frontiers in Psychology*, *9*, 1–9. https://doi.org/10.3389/fpsyg.2018.01694

Halberda, J. (2011). What is a Weber fraction? Available from: http://panamath.org/wiki/index.php?title=What_is_a_Weber_Fraction%3F [last accessed 31 January 2020].

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465. https://doi.org/10.1037/a0012682

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual Differences in non-verbal Number Acuity Correlate with Maths Achievement. *Nature*, *455*(7213), 665–668. https://doi.org/10.1038/nature07246

Hannagan, T., Nieder, A., Viswanathan, P., & Dehaene, S. (2018). A random-matrix theory of the number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170253. https://doi.org/10.1098/rstb.2017.0253

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434. https://doi.org/10.1016/j.tics.2007.09.004

Hinton, G. E., & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221–1247. https://doi.org/10.1016/j.cognition.2007.06.004

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(25), 10382–10385. https://doi.org/10.1073/pnas.0812142106

Jaeger, H., Maass, W., & Principe, J. (2007). Special issue on echo state networks and liquid state machines. *Neural Networks*, *20*(3), 287–289. https://doi.org/10.1016/j.neunet.2007.04.001

Kim, G., Jang, J., Baek, S., Song, M., & Paik, S. (2019). Spontaneous generation of innate number sense in untrained deep neural networks. *BioRxiv*, https://doi.org/10.1101/857482

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Leibovich, T., & Ansari, D. (2016). The symbol-grounding problem in numerical cognition: A review of theory, evidence, and outstanding questions. *Canadian Journal of Experimental Psychology*, *70*(1), 12–23. https://doi.org/10.1037/cep0000070

Leslie, A. M., Gelman, R., & Gallistel, C. R. (2008). The generative basis of natural number concepts. *Trends in Cognitive Sciences*, *12*(6), 213–218. https://doi.org/10.1016/j.tics.2008.03.004

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication Theory*, *84*, 486–502.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 8–45). Oxford, UK: Clarendon Press/Oxford University Press.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. *International Perspectives on Psychological Science*, *1*, 57–88.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, *4*, 503. https://doi.org/10.3389/fpsyg.2013.00503

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356. https://doi.org/10.1016/j.tics.2010.06.002

McClelland, J. L., McNaughton, B., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407.

Miller, K., Keller, J., & Stryker, M. (1989). Ocular dominance column development: Analysis and simulation. *Science*, *245*(4918), 605–615. https://doi.org/10.1126/science.2762813

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and Their Acquisition*, *1*, 1–55.

Nieder, A. (2005). Counting on neurons: The neurobiology of numerical competence. *Nature Reviews Neuroscience*, *6*(3), 177–190. https://doi.org/10.1038/nrn1626

Odic, D., Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Developmental change in the acuity of approximate number and area representations. *Developmental Psychology*, *49*(6), 1103. https://doi.org/10.1037/a0029472

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Piantadosi, S. T. (2016). A rational analysis of the approximate number system. *Psychonomic Bulletin & Review*, *23*, 877–886. https://doi.org/10.3758/s13423-015-0963-8

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., ... Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in Developmental Dyscalculia. *Cognition*, *116*(1), 33–41. https://doi.org/10.1016/j.cognition.2010.03.012

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. https://doi.org/10.1016/j.neuron.2004.10.014

Piazza, M., Pica, P., Izard, V., Spelke, E. S., & Dehaene, S. (2013). Education enhances the acuity of the nonverbal approximate number system. *Psychological Science*, *24*(6), 1037–1043. https://doi.org/10.1177/0956797612464057

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica*, *140*(1), 50–57. https://doi.org/10.1016/j.actpsy.2012.02.008

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607–614. https://doi.org/10.1111/j.1467-9280.2008.02130.x

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. https://doi.org/10.1038/14819

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: Foundations*, Vol. 1. Cambridge, MA: MIT Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Snoddy, G. (1926). Learning and stability: A psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology*, *10*(1), 1–36. https://doi.org/10.1037/h0075814

Soltész, F., Szűcs, D., & Szűcs, L. (2010). Relationships between magnitude representation, counting and memory in 4- to 7-year-old children: A developmental study. *Behavioral and Brain Functions*, *6*(13), 1–14. https://doi.org/10.1186/1744-9081-6-13

Stoianov, I., & Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. *Nature Neuroscience*, *15*(2), 194–196. https://doi.org/10.1038/nn.2996

Testolin, A., Dolfi, S., Rochus, M., & Zorzi, M. (2019). Perception of visual numerosity in humans and machines. In arXiv:1907.06996.

Testolin, A., & Zorzi, M. (2016). Probabilistic models and generative neural networks: Towards an unified framework for modeling normal and impaired neurocognitive functions. *Frontiers in Computational Neuroscience*, *10*, 73, https://doi.org/10.3389/fncom.2016.00073

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience*, *16*(9), 1493–1504. https://doi.org/10.1162/0898929042568497

Widrow, B., Greenblatt, A., Kim, Y., & Park, D. (2013). The no-prop algorithm: A new learning algorithm for multilayer neural networks. *Neural Networks*, *37*, 182–188. https://doi.org/10.1016/j.neunet.2012.09.020

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88–101. https://doi.org/10.1111/j.1467-7687.2005.00395.x

Zorzi, M., & Testolin, A. (2018). An emergentist perspective on the origin of number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170043, https://doi.org/10.1098/rstb.2017.0043

Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: A tutorial overview. *Frontiers in Psychology*, *4*, 515. https://doi.org/10.3389/fpsyg.2013.00515

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Testolin A, Zou WY, McClelland JL. Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Dev Sci*. 2020;00:e12940. https://doi.org/10.1111/desc.12940