Title

Non-numerical features fail to predict numerical performance in real-world stimuli

Authors

Sanford, Emily M

Halberda, Justin

Publication Date

2024

DOI

10.1016/j.cogdev.2023.101415

Cognitive Development 69 (2024) 101415

Contents lists available at ScienceDirect

Cognitive Development

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

Non-numerical features fail to predict numerical performance in real-world stimuli

Emily M. Sanford *, Justin Halberda

Johns Hopkins University, USA

ARTICLE INFO

Original content: Counting book illustrations (color and silhouette), feature extraction code (Original data)

Keywords:
ANS
Children's books
Naturalistic stimuli
Real-world
Counting

ABSTRACT

It has been proposed that humans use non-numerical features (such as convex hull and surface area) to estimate the number of objects in a scene. This would be an evolutionarily advantageous strategy if such features truly patterned with number in the world, but this has never been empirically tested. Here, we quantify the strength of the relationship between number and non-numerical features in two relevant image sets: the illustrations from children's counting books, and real-world photographs. We find that non-numerical features are much less predictive of the number of objects in counting books than in photographs, despite the former being specifically designed for use in teaching children about numbers. Then, across three behavioral experiments, we ask whether the stronger relationship in photographs predicts better number estimation performance in adults (N = 120) and in children (N = 94; *M* age = 7;2 years). Our experiments reveal that number estimation is easier from the counting books than the photographs, even though non-numerical features are *less* predictive of number in books. This analysis uses real-world stimuli and draws into question the claim that non-numerical features are intrinsically involved in number extraction.

The ability to perceive the number of objects in a group is shared across species and across human development (Dehaene, 1997; Feigenson & Dehaene, & Spelke, 2004). Young infants have been shown to possess numerical sensitivity (Xu & Spelke, 2000), and precision in approximate number discrimination continues to develop until about age 30 (Halberda et al., 2012). How do we accomplish this remarkable perceptual feat, especially considering that number is a very abstract concept compared to most other magnitudes (e.g., height, weight)? Some have proposed that numerical representations must be grounded in more concrete sources of evidence; for instance, number responses have been attributed to non-numerical features rather than number itself (Clayton, Gilmore & Inglis, 2015; Gebuis et al., 2016; Smets et al., 2015; Szűcs et al., 2013). Features that are often implicated as underlying "number-like" sensitivity include surface area (e.g., Gebuis et al., 2016), convex hull (Clayton et al., 2015), and density (Dakin et al., 2011; Durgin, 1995, 2008; Morgan et al., 2014), or some combination of features (e.g., Gebuis et al., 2016).

Why would such features be used by a system attempting to estimate the number of objects in a scene? One rationale is that they are ecologically useful, and have been in our habitats over the course of evolution. Indeed, it has been extensively claimed that non-numerical features pattern reliably with number in the natural world, where larger numerical magnitudes are associated with larger non-numerical magnitudes such as physical size (e.g., Abalo-Rodríguez et al., 2022; Leibovich et al., 2017; Smets et al., 2015; van Rinsveld et al., 2020). It is easy to imagine situations in which this pattern holds true. For instance, the mass of two elephants put together is much larger than one elephant alone; thus, size may correlate with number in the real world. However, it is also possible to

* Correspondence to: Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94704, USA.
E-mail address: esanford@berkeley.edu (E.M. Sanford).

https://doi.org/10.1016/j.cogdev.2023.101415
Received 1 March 2023; Received in revised form 22 November 2023; Accepted 27 December 2023

imagine situations in which this correlation would *not* be respected, as in the case of one elephant versus 10 mice. In fact, the strength of the relationship between number and non-numerical features in the natural world has yet to be empirically tested.

One previous modeling attempt (Testolin et al., 2020) used natural scenes but turned these into artificial scenes by using a bounding box around each object—turning natural scenes into scenes of white rectangles on a black background. This obscures the natural scene statistics; the real world is full of complex, non-rectangular shapes, whose area and perimeter can differ substantially from rectangles of approximately the same size. Furthermore, it has been demonstrated that object complexity drives visual attention and perception (Sun & Firestone, 2021), and thus, simplifying objects to rectangles may fail to capture visual algorithms which are refined to work on complex shapes. Another attempt used real scenes and artificial scenes constructed from these images (Odic & Oppenheimer, 2023), but did not measure the non-numerical visual features that might be key to number estimation. Here, we seek to formally characterize the extent to which non-numerical features predict number in such natural scenes—prior to any observer or task—that is, what are the correlations between number and non-numerical cues in natural scenes? Second, we seek to measure the precision of numerical decisions given these scenes to determine if human performance (adult and children) patterns with these correlations between number and non-numerical features—that is, do humans estimate number more accurately when there are non-numerical correlations to help them?

How would a system that uses non-numerical features to construct numerosity representations work? A prominent claim is that number perception involves the combination of cues from multiple continuous dimensions at once. For instance, Gebuis et al. (2016) proposed a Sensory-Integration System, which accumulates evidence from many continuous features such as surface area, convex hull, and density and combines them into one general magnitude representation. According to this theory, judging which of two ensembles contains more items would be achieved not through a direct evaluation of numerosity, but instead by determining which ensemble yields a larger total summed magnitude representation.

To gain purchase on the extent to which such features support numerical responses, we might take the tact of looking at the stimuli themselves: What are children looking at as they learn about numbers? What features are available in the environment to help guide number learning? If children learn about visual number through relationships between non-numerical features and number of objects in their visual experience, then those relationships must exist in the world for children to learn from them. To investigate whether this is the case, we began with an obvious but heretofore untapped stimulus set: children's counting books.

Counting books are an important venue through which parents introduce children to number words starting in the first year of the child's life (Goldstein et al., 2016). In general, picture books facilitate learning, and this is especially the case when the illustrations in the books are realistic (Ganea et al., 2008). Further, they specifically facilitate the development of mathematical thought and numeracy (van den Heuvel-Panhuizen et al., 2009); for instance, the simple manipulation of changing the direction in which a book is read (left to right versus right to left) influences the direction in which kids will subsequently count (Göbel et al., 2018). Previous research has investigated to what extent features of counting books are aligned with the literature on numeracy development, particularly looking at things like overall book structure (e.g., whether the numbers are listed in ascending order) and image structure (e.g., presence or absence of distractor objects; Ward et al., 2017). It has been found that, in children's counting books, the *pictures* themselves are specifically useful for eliciting numerical thought and behavior in young children (Elia et al., 2010). This is especially important considering that most counting board books are intended for very young audiences (for the books in our analysis, the average intended age range was 1–4 years old), who are only just acquiring their number word knowledge at the end of that range (Sarnecka & Carey, 2008). This motivates the question of what exactly children are learning from reading these books at such a young age.

Even though the text of counting books primarily teaches counting words and symbols, these books overwhelmingly include visual depictions of the number of items being discussed on the page, providing an explicit demonstration of what visual to associate with what digit. Therefore, counting books teach children about *visual number*: what sorts of visuals should be associated with different cardinal values. Because they are an early source of numerical exposure for children, and because each page contains illustrations intended to convey visual number information, they are a particularly interesting resource to answer the question of what relationships between number and non-numerical features children are exposed to and instructed from in early development.

We investigated the role of non-numerical features in number perception in naturalistic materials. We began with a formal analysis of the visual features available in children's counting books and how they relate to depicted number (e.g., how does an illustration of 10 balls differ from an illustration of 2 cows?). We then repeated our analysis procedure with real-world photographs of collections to see whether the evidence available in the non-numerical features of real-world images would pattern similarly to children's counting book illustrations, and whether their non-numerical magnitudes would provide a useful signal for number. Finally, we performed a series of three behavioral experiments with both adult and child participants, asking them to estimate the number of objects in scenes from each of these datasets. This allows us to evaluate the extent to which non-numerical feature information is used during number estimation. The results converge to demonstrate the importance of approximate number as an independent source of information apart from other continuous magnitudes such as area and convex hull, both in terms of the images themselves and in terms of humans' abilities to make judgments about those images.

Throughout this investigation, we focused on the numbers 1 through 10. While some have argued for two systems of representation, one for "small" sets from 1–4 (Object Tracking System, or OTS; e.g., Butterworth, 2010; Feigenson et al., 2002; Feigenson & Carey, 2003; Piazza, 2010) and another for "large" sets above 4 or 5 (Approximate Number System, or ANS; e.g., Feigenson & Dehaene, & Spelke, 2004; Piazza et al., 2004; Xu & Spelke, 2000), chunking can expand the number of objects included in set-based representations of individuals (Feigenson & Halberda, 2004; Halberda et al., 2006; vanMarle et al., 2018). However, it has been demonstrated that the ANS–the "large" number system–has representations spanning all the way down to 1 (Cordes et al., 2001; Whalen et al., 1999). Some recent modeling work has even suggested that there is only one system underlying both large and small number representation, and

this proposed unitary system is behaviorally consistent with the ANS (Cheyette & Piantadosi, 2020). Because of their extreme relevance to the early stages of number learning (Le Corre & Carey, 2007), as well as their dominance in natural scene experience (Piantadosi, 2016), ensembles containing 1 to 10 objects struck us as particularly important to study.

1. Experiment 1

In Experiment 1, we evaluate whether non-numerical continuous features predict the number of objects displayed in the illustrations of children's counting books.

2. Method

2.1. Book preparation

We purchased children's counting books from the Amazon marketplace (see Fig. 1), with each book costing between $3 and $10. We selected only books that were intended to teach children about numbers and counting, and that contained explicit instruction of the numbers one through ten, as well as illustrations of ensembles containing each of those values (one book only contained depictions of one through five). We purchased the top 60 books on the Amazon marketplace that appeared to meet the above criteria, and only included those that actually met the criteria in the actual sample. Some authors had multiple books available; in the final analysis, we included 49 books that were illustrated by 42 different people, with at most five books coming from one illustrator (see Supplemental Materials). Seven of the books were soft cover books and the remainder were board books of various sizes.

Once the books were selected, the pages of each book were manually scanned, with all pages from a given book saved to an image of the same size. The scanned images were edited to prepare for feature extraction using a combination of MS Paint and Adobe Photoshop. The purpose of this editing was to remove all irrelevant background and text details so that only the countable objects remained. When images contained overlapping objects, we created a two-pixel boundary between the objects to make them separately countable (16% of images). The images were then converted to black and white images using a custom Python script, where the countable objects were colored with black pixels and the rest of the image was filled in with white pixels.

2.2. Feature extraction

The black and white images were processed using a custom feature extraction algorithm created in Python (see Fig. 2). The algorithm extracted three group-level features: total surface area (number of black pixels in the image), total object perimeter (number of black pixels with at least one neighboring white pixel), and convex hull (the area contained within the perimeter around the entire group, calculated using the convex hull function from Python's SciPy package; Jones et al., 2001). This procedure for calculating convex hull means that there is a non-zero value for convex hull for even one-item images, since it computes the convex hull around all *pixels* rather than the convex hull around object centroids. Because ensemble statistics (Sweeny et al., 2015), in addition to total values, may be important for generating numerical estimates, the group-level features and the number of items per page were then used to calculate the average features of each image: average surface area per item (total surface area divided by number), average perimeter per item (total perimeter divided by number), and average convex hull per item (convex hull divided by number). Average features appear to be estimated psychologically *without* attending to individual objects and so may not require knowledge of the number of

Fig. 1. Example illustrations from children's counting books.

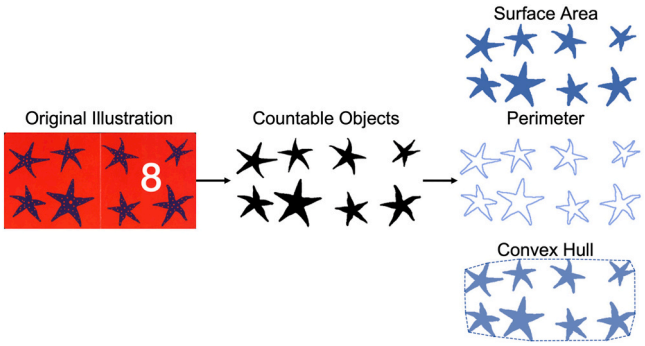

Fig. 2. Example feature extraction from children's counting book illustrations. Group-level features represent all pixels from the entire image (e.g., all blue pixels in the Surface Area image); Average-level features represent the average number of pixels per object in the image (e.g., the number of pixels in one starfish in the Surface Area image).

items on the part of the observer (Alvarez, 2011; Ariely, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2012; Ward et al., 2016). This resulted in a total of six non-numerical feature values per image (Total Surface Area, Convex Hull, Total Perimeter, Average Surface Area, Average Convex Hull, and Average Perimeter).

3. Results

3.1. Data cleaning

For some books, there was more than one illustration for a given number (e.g., one book contained illustrations of "two stars" as well as "two rocket ships" for the number two), and in those cases both images were included in analysis. With these duplicates, we extracted feature information from a total of 522 illustrations.

Once we had acquired feature information for all dimensions of interest from each illustration, we used a trimming procedure to remove outliers with respect to each dimension. We did this because we were most interested in how the typical features available in such materials related to their numerical content. We utilized a method involving the Median Absolute Deviation because it is more robust to non-normal data, and removed datapoints that were more than three times the MAD away from the median value (Leys et al., 2013). We did this separately for each feature (e.g., surface area, convex hull, etc.) for each numerical value (1−10). Images that were identified as outliers for any feature were removed from all subsequent analyses. This procedure resulted in the removal of 52 images (≈10% of images, 4 to 8 images per number). Therefore, all subsequent analyses are performed on data from the remaining 470 illustrations, with between 44 and 49 images per number.

It is important to note that, for both Experiments 1 and 2, the removal of outliers did not impact the results of our analyses. Without outliers removed, the relationship between number and non-numerical features tended to decrease by a small amount (with r^2 values changing by at most .063). The largest change between the full and trimmed datasets was the group-level model in the counting books dataset, whose R^2 value decreased by .05 with outliers included. None of the statistical tests changed with respect to their significance (e.g., all *p*s <.001 remained so).

3.2. Individual feature correlations

For our first analysis, we evaluated the strength of the linear relationship between number and each measured feature using correlations (see Table 1). We found significant linear relationships with all features, with effect sizes ranging from small (r^2 ≈.01: Total Surface Area and Convex Hull) to medium (.09 < r^2 <.25: Total Perimeter, Average Surface Area, Average Convex Hull, and Average Perimeter; Cohen, 1992). Qualitatively, the magnitude of the relationship between number and the average-level features (such as average surface area) was stronger (average r^2 =.187) than the magnitude of the relationship between number and the group-level features (average r^2 =.092), although this difference was not statistically evaluated. Nonetheless, it is worth noting since the group-level features are much more frequently discussed in numerical perception literature (e.g., Clayton & Gilmore, 2015).

Table 1
Correlations between number and each non-numerical feature across the illustrations of children's counting books.

	Total Surface Area	Convex Hull	Total Perimeter	Average Surface Area	Average Convex Hull	Average Perimeter
r^2	.011 *	.075 ***	.190 ***	.220 ***	.159 ***	.182 ***

Note. * $p < .05$, *** $p < .001$ (Holm correction)

3.3. Combination models

Next, we investigated whether combinations of evidence from multiple features could provide a stronger relationship between number and non-numerical continuous feature signals. We used two different methods to combine information from different features, resulting in three models predicting number from the non-numerical features in the images.

The first method we employed was a Principal Component Regression approach (PCR; Liu et al., 2003). Because there is high collinearity between the non-numerical features, it is not recommended to directly include them as separate regressors in a linear regression model (Næs & Mevik, 2001). Principal Component Analysis (PCA) is a valuable statistical technique for dimensionality reduction, meaning that it allows us to reduce the number of variables in a dataset while retaining as much information as possible. The variables created through this technique are known as Principal Components (PCs) and can be used to summarize the primary directions of variance in the dataset. Importantly, the PCs are defined to be orthogonal to one another, so multicollinearity ceases to be an issue. We can then use the PCs as regressors in a linear model, and this allows us to evaluate how much variance in the outcome variable (number) is accounted for by the entire suite of variables inputted to the PCA as a whole. Importantly, the PCs are simply linear combinations of the variables inputted into the analysis. For a given set of input variables, the full model (i.e., the one containing all PCs as predictors) is exactly identical to a model containing the original input variable set in terms of variance explained and outcome variable (Number) predictions.

A primary limitation of this technique is that we will not be able to determine precisely *which* of the features included in the model is responsible for the most variance in number. Notably, this is also not possible when directly inputting collinear predictors into a regression model. Therefore, we will restrict our interpretation to whether the *group of features altogether* accounts for substantial variance in number. This is compatible with the literature on the role of non-numerical features in numerical processing; if people are using non-numerical features to generate their number estimates, it is not usually claimed that people use fully-formed representations of independent features such as surface area and perimeter; instead, they might use a general magnitude composite or a linear combination of information from different dimensions (e.g., Gebuis et al., 2016; Walsh, 2003). We can think of the components generated by our PCA as candidate composite non-numerical feature representations.

The general procedure was the same for each multi-feature model that we built. First, we performed a PCA over a particular set of features using the 'prcomp' function in R (R Core Team, 2013). In a PCA, the primary directions of variance in the dataset are captured by independent vectors, which are then used as the regressors in our models. We used this to extract the information contained within a given group of non-numerical features (such that we can evaluate how well that group *as a whole* predicts number), while ensuring that the resulting regressors are orthogonal to one another.

We then progressively added the components to a linear model predicting number, in descending order based on the raw correlation between each component and number. We evaluated at each step whether the additional component explained significantly more variance in number than the previous model using ANOVA model comparisons. The final selected model for each feature group was the one with the most included features that still significantly outperformed the previous model. We chose a stepwise modeling approach because our goal was to use an analytic technique to determine the most parsimonious model–that is, the model with the fewest regressors that captured the most variance. By adding the regressors in the order of most to least relationship with number, we ensure that the final model will be maximally efficient. We did this in the hopes of maximizing psychological plausibility by minimizing the number of variables used to generate the prediction. However, note that the full model (containing all PCs) was essentially identical to the presented model in terms of its numerical predictions in every case, and our claims would remain the same whether we used a subset model or the full model.

Because our group and average-level features are transformations of one another as a direct function of number (average surface area = total surface area / number of objects), including both in the PCA would be akin to including number as both the independent and dependent variables of the regression; therefore, we only ever included them in separate models to assess the extent to which *independent* non-numerical features predict variance in number. To summarize, we created two PCR models, one predicting number from group-level features (Total Surface Area, Convex Hull, and Total Perimeter) and another predicting number from average-level features (Average Surface Area, Average Convex Hull, and Average Perimeter).

For our second modeling approach (and our third model overall), we used a specific algebraic combination that has been implicated extensively in the literature on number perception: density (DeWind et al., 2015). Although the exact definition of density varies (e.g., ratios of spatial frequency channels; Dakin et al., 2011), here we define density in terms of other implicated non-numerical features, such that density is equal to the total surface area divided by the convex hull, consistent with previous modeling of the ANS (DeWind et al., 2015). This density measure captures the amount of area over which the group is spread (convex hull) that is occupied by the countable objects themselves (surface area). This ratio would increase if the objects were packed more densely, as might be expected with an increasing number of objects depicted within a constrained image size. For the density model, we calculated density (surface area divided by convex hull) for each illustration, then performed a linear regression predicting number from density. Here we treat density as a separate modeling approach because it is not directly measured from the images, but is instead derived from the other directly-measured features. Therefore, we consider it a composite feature. We also want to differentiate it from the type of density that is often discussed in the numerical perception literature, texture density, which is purported to be directly perceived and to influence numerosity perception, but only in much more crowded displays with smaller objects than those that we are using here (e.g., Dakin et al., 2011; Durgin, 1995, 2008; Morgan et al., 2014).

To compare the non-nested models across feature types (group-level versus average-level features versus density), we computed both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), as these measures of goodness of fit can be compared even when the models are not nested with one another (Vrieze, 2012). Both metrics take into account the amount of variance

in the dependent variable accounted for by the model as well as the number of parameters in the model. The BIC penalizes the model more for additional parameters and is therefore considered a more conservative statistic than the AIC. For both the AIC and the BIC, the model with the smaller value is considered to provide a better fit to the data.

3.4. Model 1: group-level features

For our first model, we performed our PCR procedure over the three independently-extracted group-level features, total surface area, total perimeter, and convex hull (see Fig. 3A). In the PCA, the first component (PC1), explained the majority of the variance in the three features (85.95%) but correlated more weakly with number (r^2 =.085) than did PC2 (9.79%, r^2 =.191). The third component (PC3) both explained the least variance in the non-numerical features (4.26%) and correlated the weakest with number (r^2 <.001). Therefore, the components were added to the stepwise regression in the following order: PC2, then PC1, then finally PC3. Loadings are shown in Table 2.

The model fits for each step of the regression are presented in Table 3. The first model, using only PC2, explained 19.0% of the variance in number, while the second model (which included PC2 and PC1) explained 27.4% of the variance in number, and this improvement was significant, $F(1, 467) = 55.15$, $p < .001$. The addition of the third component did not explain significantly more variance in number, $p = .932$, and therefore we select Model 2 as the best group-level feature model.

Next, we evaluated how accurately the number estimation model built from these non-numerical features represented number. Once the regression weights are set across the entire dataset, we can calculate the model's predicted value ($\hat{y}$) for how many objects should be present in the image, given that image's values for the PCs. We can do this for every image, then compare those values to the true number of objects that are actually present in the images (see Fig. 4A). We found that the average predicted number for each true number ranged from 4.06 for N = 1 to 6.66 for N = 10.

Using one-sample t-tests, we evaluated whether the model's average estimate was significantly different from the true number of objects depicted in each illustration. For all numbers besides N = 5 and N = 6, whose predicted values did not differ significantly from true values ($ps > .151$), we found that the model's estimate was significantly different from the true number depicted, $ts > 5.16$, $ps < .001$. In summary, the group-level non-numerical features model did a poor job recovering the true number of objects depicted

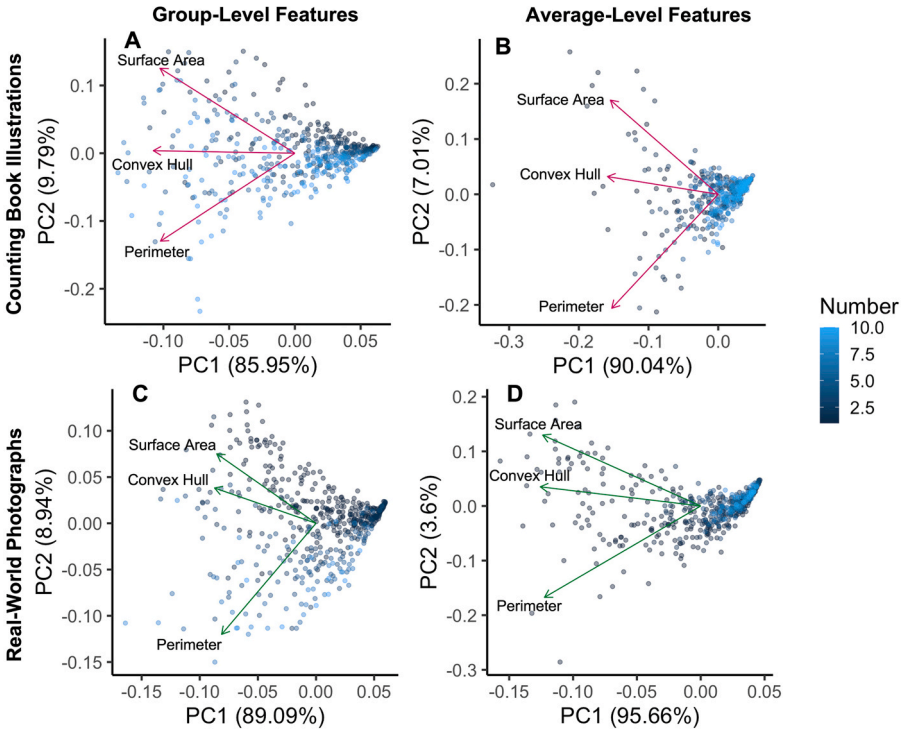

Fig. 3. Principal Component Analysis for non-numerical features in Experiment 1 (children's counting books, top row) and Experiment 2 (photographs, bottom row). Principal Components (PCs) capture the strongest information represented by the features from which they are derived. In the plot, the PCs represent the main directions of variance in those features. Each point represents one illustration/photograph's value for the first two PCs, and the arrows represent where the original features lie relative to the new coordinate space. For instance, in the group-level feature analysis of counting books (A), Convex Hull is nearly parallel to PC1, indicating that the information captured by PC1 is highly similar to information contained within the Convex Hull variable. Points are colored by the number of objects in the illustration/photograph. For example, the vertical color gradient across illustrations in Panel A (from light blue at the bottom to dark gray at the top) indicates that PC2 captured more variance in number than PC1.

Table 2
PCA loadings for group-level features; All three features had nearly equal negative loadings on PC1; Total Surface Area positively loaded on PC2 while Total Perimeter negatively loaded on it; and Convex Hull negatively loaded on PC3.

	PC1	PC2	PC3
Total Surface Area	-.57	.69	.44
Convex Hull	-.60	.02	-.80
Total Perimeter	-.57	-.72	.40

Table 3
Stepwise PCR models predicting number in counting book illustrations using group-level features.

	Model 1		Model 2		Model 3	
Regressor	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*
PC2	-2.31	0.22	-2.31	0.21	-2.31	0.21
PC1			-0.52	0.07	-0.52	0.07
PC3					0.03	0.32
Adjusted R^2	.190		.274		.272	
F for ΔR^2			55.15 ***		.01	

Note. *** $p < .001$

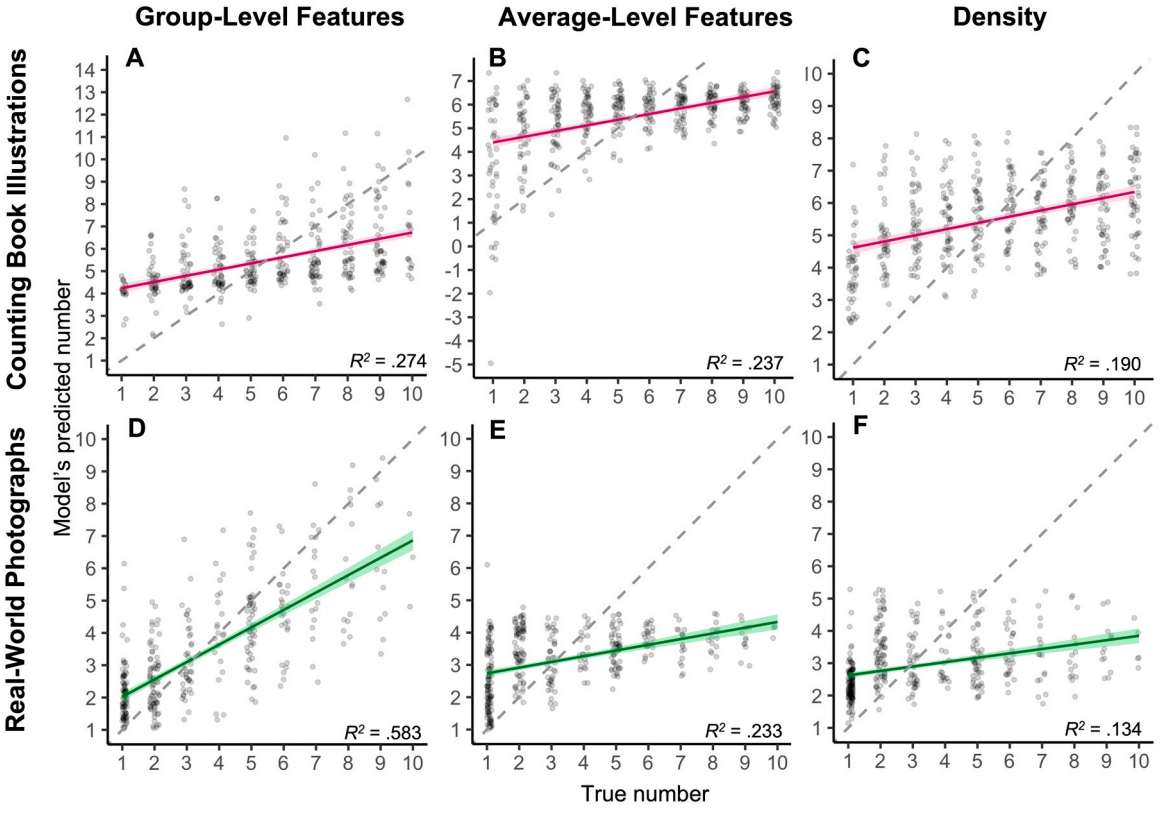

Fig. 4. Relationship between true number (jittered) and models' predicted number of objects ($\hat{y}$) in Experiment 1 (counting book illustrations, top row) and Experiment 2 (real-world photographs, bottom row). The gray dashed line depicts a perfect prediction. Most of the models fail to capture number in a variety of ways: for example, in (A), the slope is too shallow (the model provides an average guess of ≈4–6 objects for all images, regardless of their true number), and in (F), the model accounts for very little variance in number (R^2 =.134). The best performing model (D) has the steepest slope (meaning that its estimates are the most accurate) and accounts for the most variability in number. *Note.* Gray dashed line depicts a perfect prediction.

across the illustrations.

3.5. Model 2: average object-level features

For our second model, we performed our PCR procedure over the three independently-extracted average-level features: average surface area, average perimeter, and average convex hull (see Fig. 3B). We expected this model might outperform the previous one because the average-level features of counting book illustrations individually correlated more strongly with number than did the group-level features. We performed a PCA over these three features, then predicted the number of objects from the resulting components.

In the PCA, the first component (PC1) explained a vast majority of the variance in the three features (90.04%) and also correlated most strongly with number (r^2 =.206). The third component (PC3) explained the least variance in the non-numerical features (2.95%) but correlated more strongly with number (r^2 =.034) than PC2 (7.01%, r^2 =.002). Therefore, the components were added to the stepwise regression in the following order: PC1, then PC3, then finally PC2 (for the component loadings, see Table 4).

The model fits for each regression step are presented in Table 5. The first model, using only PC1, explained 20.5% of the variance in number, while the second model (which included PC1 and PC3) explained 23.7% of the variance in number, and this improvement was significant, $F(1, 467) = 20.81, p < .001$. The addition of PC2 in the third model did not explain significantly more variance in number, $p = .331$, and therefore we select Model 2 as the best average-level feature model.

Next, we once again evaluated the accuracy with which this model represented the number of objects across the illustrations (comparing y and $\hat{y}$; see Fig. 4B). We found that the average predicted number for each true number ranged from 3.38 for N = 1 to 6.24 for N = 10.

We again used one-sample t-tests to evaluate whether the model's average estimate was significantly different from the true number of objects depicted in each illustration. Estimates of N = 6 did not significantly differ from the true value, $p = .157$, but we found that the model's estimate was significantly different from the true number depicted for every other number, $ts > 7.16, ps < .001$.

Although the average-level features individually patterned more strongly with number than did the group-level features, possibly due to the strong amount of overlapping variance between them, this multiple-feature model failed to outperform the prior model.

3.6. Model 3: density

For the third combination model, we evaluated to what extent pixel density (surface area divided by convex hull) predicted the number of objects in the images. We calculated a density value for each illustration, then performed a linear regression predicting number from density. We found that density significantly predicted number, $B = -6.03$, $SE\ B = 0.57$, $R^2 = .190$. When we evaluated the extent to which the model accurately represented the number of objects, we found that the average predicted values ranged from 3.84 for N = 1 to 6.22 for N = 10 (see Fig. 4C; note that because our calculation of density derives from total surface area divided by convex hull, there is variation in density even for set size 1). Aside from estimates of N = 6 which were not significantly different from the true value, $p = .505$, the average model estimates were significantly different from the true value for all other numbers, $ts > 3.96$, $ps < .001$. Somewhat surprisingly, the density model performed worse than either of the linear feature combination models.

3.7. Model comparison across feature types

Now that we have our best-performing models for each feature type (group-level features, average-level features, and density), we sought to determine which features best predicted number overall using AIC and BIC, where a smaller value corresponds to a better model fit. The group-level features model provided the best fit (AIC = 843.12; BIC = 2195.54), followed by the average-level features model (AIC = 866.17; BIC = 2218.58), and the density model performed the worst overall (AIC = 893.38; BIC = 2241.65). Therefore, we conclude that the group-level features model, which accounted for only 27.4% of variance in number, nonetheless performed the best of any of the continuous feature models we tested.

4. Discussion

We evaluated the extent to which non-numerical features could predict the number of objects across the illustrations of children's counting books. We found that there were at best weak relationships between number and non-numerical visual features, with the group-level feature model outperforming the other tested models. This is surprising, as we would have expected strong relationships

Table 4
PCA loadings for the average-level features; All three features had negative loadings of similar magnitudes on PC1; Average Surface Area positively loaded on PC2, while Average Perimeter negatively loaded on it; and Average Convex Hull was the primary source of variance (negatively) captured by PC3.

	PC1	PC2	PC3
Average Surface Area	-.57	.63	-.52
Average Convex Hull	-.59	.12	-.80
Average Perimeter	-.57	-.77	-.30

Table 5
Stepwise PCR models predicting number in counting book illustrations using average object-level features.

	Model 1		Model 2		Model 3	
Regressor	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*
PC1	.792	0.07	.792	0.07	.792	0.07
PC3			1.78	0.39	1.78	0.39
PC2					-.245	0.25
Adjusted R^2	.205		.237		.237	
F for ΔR^2			20.91 ***		.946	

Note. *** $p < .001$

between number and non-numerical features in pedagogical materials about number, if said features are useful for number perception. However, it is possible that illustrators are merely recapitulating the statistics of their day-to-day visual experience in these texts, and that this relationship is not as strong in the world as has been previously guessed. We investigate this empirically in Experiment 2.

5. Experiment 2

Here, we investigated to what extent continuous, non-numerical features pattern with number in the natural world. That is, can we find evidence for shared variability between number and non-numerical features across a database of photographs of real-world scenes?

6. Method

6.1. Image selection

We selected images from the Microsoft Common Objects in Context (MS COCO) database (Lin et al., 2014). This database contains thousands of images including indoor and outdoor scenes, natural and man-made objects, and a variety of animals. Importantly, this database comes with outlines for each object category in each picture, which were previously drawn by human participants. Therefore, for each image in the database, we can convert it to a black and white image, where black pixels correspond to the countable objects and the rest of the image is made of white pixels (essentially eliminating the background). From this image, we can easily extract continuous feature information using the extraction algorithm described in Experiment 1. Since the outlines were not drawn specifically for this purpose (unlike in Experiment 1), the amount of space between the outlines of overlapping objects was not controlled.

Although this dataset contains segmentation information for many different types of objects, we chose to focus on three targets: birds, sheep, and motorcycles (see Fig. 5). These categories were chosen because previous work had already estimated the number of each of these groups in the database (Rajan et al., 2019), thereby giving us a ready value with which we could compare the non-numerical feature statistics (although we also manually counted the number of objects in all images included in analyses to check for any count errors). Further, these three categories span a few useful dimensions: there are both man-made and natural items, they vary in real-world size, and they are often photographed from various distances (e.g., a close-up photograph of a motorcycle may be easier to achieve than a close-up of a sheep). Throughout the database, the collections of items are photographed from a variety of distances—creating some variability. However, by focusing on just three categories, it may be that we will find a stronger relationship between number and non-numerical features than in children's counting books—where we relied on many different books. Thus, our case study here presents what might be considered a best-case scenario for finding a relationship between numerical and non-numerical features in real-world images.

From these three categories, there were 1049 images available in the Val2014 dataset in the 1–10 object range (chosen for consistency with the counting book analysis). Of those 1049 images, nearly half (512) depicted only one instance of a target object. Therefore, we decided to select a subset of this image set to have a more balanced numerical distribution for analysis. In our subsample, we included all images depicting N = 5 through N = 10 targets (130 images), and randomly sampled additional images from N = 1 through N = 4 to have an equivalent number of images as in the counting book analyses (522 images total).

Fig. 5. Example photographs from the categories bird, motorcycle, and sheep.

To ensure that we did not select a non-representative portion of the dataset, we performed this sampling 1000 times and averaged them to compare with a representative individual sample. We also performed all analyses on the full image set as well as on average feature values for each number, and all of these results were highly similar, so for clarity we will only report the single subset analysis here.

6.2. Feature extraction

We used the same feature extraction algorithm as in Experiment 1, and again this resulted in a set of 6 non-numerical features per photograph (total and average surface area, perimeter, and convex hull; see Fig. 6).

7. Results

7.1. Data cleaning

As in the counting book analysis, we trimmed the dataset to remove images that were greater than three MADs from the median per number on any non-numerical feature (surface area, convex hull, perimeter, or their average-object level equivalents; Leys et al., 2013). This removed 23 images, resulting in 499 images included for analysis.

7.2. Individual feature correlations

For our first analysis, we evaluated the strength of the linear relationship between number and each measured feature (see Table 6). We found significant linear relationships with all features. The effect sizes ranged from small ($r^2 \approx .01$: Total Surface Area and Convex Hull) to medium ($.09 < r^2 < .25$: Average Surface Area, Average Convex Hull, and Average Perimeter) to large ($r^2 > .25$: Total Perimeter; Cohen, 1992). In general, we found that the average object-level features were more correlated with number than the group-level features, with the notable exception of total perimeter, which had the strongest relationship (although, again, these differences were not tested statistically).

7.3. Combination models

Next, we again turned to combination models (group-level features, average level features, and density) to investigate whether the composite of multiple features can explain additional variance in number. We use the same PCR procedure and density model as described in Experiment 1.

7.3.1. Model 1: group-level features

For the first model, we performed our PCR procedure over the three independently extracted group-level features: total surface area, total perimeter, and convex hull (see Fig. 3C). In the PCA, the first component (PC1) explained the majority of the variance in the three features (89.09%) but correlated more weakly with number ($r^2 = .080$) than did PC2 (8.94%, $r^2 = .494$). The third component (PC3) both explained the least variance in the non-numerical features (1.98%) and correlated the weakest with number ($r^2 = .011$). Therefore, the components were added to the stepwise regression in the following order: PC2, then PC1, then finally PC3. Loadings are shown in Table 7.

The model fits for each step of the regression are presented in Table 8. With only one component (PC2), the first model explained

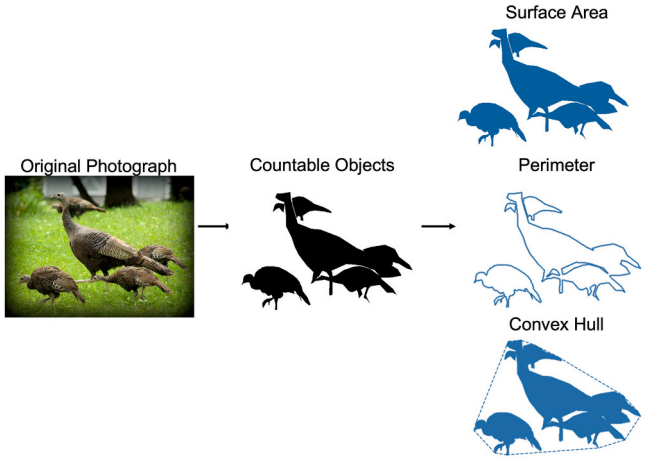

Fig. 6. Example feature extraction from photographs.

Table 6
Correlations between number and each non-numerical feature across photographs.

	Total Surface Area	Convex Hull	Total Perimeter	Average Surface Area	Average Convex Hull	Average Perimeter
r^2	.004	.040 ***	.302 ***	.182 ***	.179 ***	.223 ***

Note. *** $p < .001$

Table 7
PCA loadings for group-level features in real world photographs; all three features negatively loaded on PC1; Total Surface Area positively loaded on PC2, while Total Perimeter negatively loaded on it; and Total Surface area positively loaded on PC3, while Convex Hull negatively loaded on it.

	PC1	PC2	PC3
Total Surface Area	-.58	.51	.63
Convex Hull	-.60	.26	-.76
Total Perimeter	-.55	-.82	.15

Table 8
Stepwise PCR models predicting number in real-world photographs using group-level features.

	Model 1		Model 2		Model 3	
Regressor	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*
PC2	-3.29	0.15	-3.29	0.14	-3.29	0.14
PC1			-0.42	0.04	-0.42	0.04
PC3					-1.04	0.29
Adjusted R^2	.493		.572		.583	
F for ΔR^2			92.97 ***		13.02 ***	

Note. *** $p < .001$

49.3% of the variance in number. The second model, which included PC2 and PC1, explained 57.2% of the variance in number, and this improvement was significant, $F(1, 496) = 92.97$, $p < .001$. With the addition of the third component, the third model explained 58.25% of variance in number, and this also explained significantly more variance, $F(1, 495) = 13.02$, $p < .001$. Therefore, we select Model 3 as the best group-level feature model.

Next, we also evaluated the accuracy of this model's number estimates (y versus $\hat{y}$; see Fig. 4D). We found that the average predicted number for each true number ranged from the smallest at 1.66 for N = 1 to the largest at 6.18 for N = 9.

Using one-sample t-tests, we evaluated whether the model's average estimate was significantly different from the true number of objects depicted in each photograph. For all numbers besides N = 3 and N = 4, for which we did not find evidence that they were represented inaccurately ($ps > .710$), we found that the model's estimate was significantly different from the true number depicted, $ts > 3.38$, $ps < .002$. In terms of the overall pattern and total amount of accounted for variance, this model did a relatively good job capturing numerical variability.

7.3.2. Model 2: average object-level features

For our second model, we performed our PCR procedure over the three independently extracted average-level features: average surface area, average perimeter, and average convex hull. As we saw before, while average surface area and average convex hull were each independently more related to number than their group-level counterparts, average perimeter was much less strongly related than total perimeter, which we predicted could lead to the average-level model performing worse than the previous model.

Once again, we performed a PCA over these three features, then predicted the number of objects from the resulting components (see Fig. 3D). In the PCA, the first component (PC1) explained a vast majority of the variance in the three features (95.66%) and also correlated most strongly with number (r^2 =.203). The second component (PC2) explained the next most variance in non-numerical features (3.60%) but correlated slightly less strongly with number (r^2 =.015) than PC3 (.74%, r^2 =.020). Therefore, the components were added to the stepwise regression in the following order: PC1, then PC3, then PC2. Loadings are shown in Table 9.

Table 9
PCA loadings for average-level features in real world photographs; all three features negatively loaded on PC1; Total Surface Area positively loaded on PC2, while Total Perimeter negatively loaded on it; and Total Surface area negatively loaded on PC3, while Convex Hull positively loaded on it.

	PC1	PC2	PC3
Average Surface Area	-.58	.60	-.55
Average Convex Hull	-.59	.16	.79
Average Perimeter	-.57	-.78	-.26

Table 10
Stepwise PCR models predicting number in real-world photographs using average object-level features.

	Model 1		Model 2		Model 3	
Regressor	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*
PC1	0.64	0.06	0.64	0.06	0.64	0.06
PC3			2.28	0.64	2.28	0.64
PC2					0.91	0.29
Adjusted R^2	.201		.219		.233	
F for ΔR^2			12.58 ***		9.93 **	

Note. ** $p < .01$,*** $p < .001$

The model fits for each regression step are presented in Table 10. The first model explained 20.1% of the variance in number, while the second model explained 21.9% of the variance in number, and this improvement was significant, $F(1, 496) = 12.58$, $p < .001$. The addition of PC2 in the third model increased the accounted-for variance to 23.3%, and this improvement was again significant, $F(1, 495) = 9.93$, $p = .002$. Therefore, we select Model 3 as the best average-level feature model.

Next, we once again evaluated the accuracy with which this model represented the number of objects across the photographs (comparing y and $\hat{y}$; see Fig. 4E). We found that the average predicted number for each true number ranged from 2.16 for N = 1 to 3.83 for N = 10.

One-sample t-tests were used to evaluate whether the model's average estimate was significantly different from the true number of objects depicted in each photograph. We found that the model's estimate for 3 was not significantly different from 3, $p = .391$, while the estimates significantly differed from the true number depicted for every other number, $ts > 5.68$, $ps < .001$. Overall, this model performed worse predicting number across photographs than did the group-level feature model.

7.3.3. Model 3: density

For our third feature combination model, we once again calculated a typical measure of density used in numerical cognition (total surface area divided by convex hull) for each illustration, and predicted number from this value. We found that density significantly predicted number but accounted for a smaller amount of variance compared to the previous models, $B = -4.19$, $SE\ B = 0.47$, $R^2 = .134$. When we evaluated the extent to which the model accurately represented the number of objects, we found that the average predicted values ranged from the smallest at 2.39 for N = 1 to the largest at 3.51 for N = 9 (see Fig. 4F). We found that the model's representation of N = 3 was not significantly different from 3, $p = .647$, while the average model estimates were significantly different from the true value for all other numbers, $ts > 4.17$, $ps < .001$. As in the children's counting book illustrations in Experiment 1, the density model performed worse than the linear combination models.

7.4. Model comparison across feature types

Next, we once again sought to determine which features (group versus average versus density) best predicted number overall using AIC and BIC. As in Experiment 1, the group-level features model provided the best fit (AIC = 450.68; BIC = 1889.85), followed by the average-level features model (AIC = 754.11; BIC = 2193.27), and the density model performed the worst overall (AIC = 812.84; BIC = 2243.58). Therefore, we once again conclude that the group-level features model performed best of any of the continuous feature models we tested.

7.5. Model comparison across image sets

In Experiment 1, we found that the best performing model predicting number from combinations of non-numerical features was the group-level features model, which accounted for 27.4% of variance in number. In this experiment, examining the same relationship in real-world photographs, the best performing model was also the group-level features model, which accounted for 58.3% of the variance in number. Our next question was whether this difference in variance is significantly different between image sets. To test this, we compared the AIC and BIC for the best-performing model from each image set. The real-world photographs model (AIC = 450.68; BIC = 1889.85) vastly outperformed the counting books model (AIC = 843.12; BIC = 2195.54) according to both of these metrics. Additionally, the average level real-world photograph model (AIC = 754.11; BIC = 2193.27) also outperformed the average-level counting books model (AIC = 866.17; BIC = 2218.58), and the photograph density model (AIC = 812.84; BIC = 2243.58) outperformed the counting books density model (AIC = 893.38; BIC = 2241.65) according to AIC (the fit was similar when comparing with BIC). This indicates that, across feature types, continuous features better predict the number of objects in real-world images than in counting books.

8. Discussion

In Experiment 2, we investigated whether non-numerical features could provide evidence for number in naturalistic photographs. We found that non-numerical features in photographs account for a significant amount of variability in number, with the strongest performing model (using group-level features) accounting for 58.3% of the variance in number across pictures. This is significantly

more signal than accounted for by the average-level features and density models, and is also significantly better than any of the models predicting number in children's counting book illustrations (maximum $R^2 = 27.4\%$).

Group-level features such as surface area and convex hull have been strongly implicated as potential sources of information for the extraction of number from visual scenes (Clayton & Gilmore, 2015; Gebuis et al., 2016), and it has been previously asserted that these features tend to pattern with number in the real world (e.g., Abalo-Rodríguez et al., 2022; Leibovich et al., 2017; Smets et al., 2015; van Rinsveld et al., 2020). Here we provide the first empirical evidence that they *do* pattern with number in the world and could therefore be expected to pattern with number in our day-to-day visual experience. Notably, such features appear to do better in combination than in isolation, which is consistent with sensory integration and generalized magnitude theories (Gebuis et al., 2016; Leibovich et al., 2017).

Interestingly, these feature relationships were much lower in counting books than they were in photographs, despite the fact that the purpose of such books is to teach children about visual number. This is surprising if non-numerical features are obligatory for number perception, as we would have expected to find strong relationships between non-numerical features and number in materials explicitly designed for teaching children about visual number.

In light of this surprising result, two possible interpretations emerge. First is that counting books are poorly created, not depicting important relationships that are vital to extracting number from visual scenes. Second is that the lack of these relationships in counting book illustrations may indicate that such features are not as important for numerical perception as previously hypothesized. That is, perhaps books do not depict this relationship because their inclusion is not obligatory nor helpful for perceiving number (despite existing in natural visual experience).

To provide some evidence to distinguish between these possibilities, we next asked whether the weaker relationship between non-numerical features and number in counting book illustrations compared to photographs would translate to impaired *perception* of number in those images. In Experiment 3, we looked at adult estimation performance on stimuli drawn from each of these image sets.

9. Experiment 3

In this experiment, we ask human adults to estimate the number of objects in photographs or counting book illustrations. If non-numerical features are necessary for the extraction of approximate number from a visual scene, we would expect people to be more precise when estimating the number of objects in photographs (where such relationships are relatively strong) compared to counting book illustrations (where such relationships are relatively weak). If such features are *not* necessary nor helpful for number perception, we might expect to see no difference or even the opposite effect (e.g., relationships with non-numerical features causing distraction from the target feature).

10. Method

10.1. Participants

A total of 56 people (N = 28 in each condition) participated in this experiment. Participants were recruited from Prolific and were paid at a rate of approximately $14 per hour for their participation. Demographic information about participants was not collected. All experiments were approved by the Homewood Institutional Review Board at Johns Hopkins University. Informed consent was obtained prior to the start of the study, and participants were debriefed about the purpose of the study after it was concluded.

10.2. Materials

We selected a subset of 80 images each from the full sets of photographs and counting book illustrations to serve as experimental stimuli. Those 80 images were evenly divided amongst the numbers 1 through 10 (8 of each). Importantly, the stimulus sets were intentionally selected to ensure that the relationship between number and non-numerical features in the stimulus sets themselves remained significantly higher in the photograph condition (R^2 =.417, AIC = 129.63, BIC = 368.19) than in the counting book illustration condition (R^2 =.149, AIC = 159.83, BIC = 398.38).

10.3. Procedure

Participants were tested individually on their own devices over the internet, so factors such as display size and viewing distance were not controlled. Prior to the experiment, participants were randomly assigned to either the counting book or photograph condition (i.e., no one saw stimuli from both datasets). We employed a between-subjects design in order to keep the experiment relatively short and consistent for participants (i.e., to avoid switching between stimulus sets with different feature profiles). As we were primarily interested in whether the stimulus set influenced performance and participants were randomly assigned to the two conditions, this still allowed us to answer our primary question of interest.

Stimuli were presented in a different random order for each participant. Participants were informed that they would be estimating the number of different objects in images or illustrations. Each trial started with one of three text labels that was displayed for 1000 ms at the center of the screen. For the photographs, there were three possible labels: MOTORCYCLES, BIRDS, and SHEEP. There were 59 unique labels for the counting book illustrations, including BERRIES, BUILDINGS, and BASEBALL PLAYERS. After the display of the text label, the stimulus image was displayed for 800 ms before disappearing. Then, a slider bar ranging from 0 to 12 appeared with a

random starting position; the participant could manipulate the value selected on the slider using arrow keys or the mouse. The slider remained on the screen until the participant pressed the Spacebar to submit their answer, at which point the next trial began. Responses as well as response times were recorded. The experiment took approximately 10 min to complete.

10.4. Modeling

To assess the precision with which participants were estimating the number of objects in the images, we used PsiMLE (Odic et al., 2016), which allows for the simultaneous estimation of multiple parameters of interest using Maximum Likelihood Estimation (MLE). Here we use a power curve to model responses (e.g., Stevens, 1961), but PsiMLE also includes linear and logarithmic models, and with this data they produced nearly identical results.

According to the power model, estimation responses are drawn from a Gaussian distribution with a mean of N^{β} and a standard deviation of $N^{\beta} * \sigma$, where N is the true number, β corresponds to how much the representation is compressed ($\beta < 1$) or expanded ($\beta > 1$) relative to the true value, and σ is an index of internal variability or precision (smaller σ corresponds to a more precise representation). Each participant is assumed to have a stable value for each parameter that underlies their responses across different values of N. Using PsiMLE, we estimate one β and σ for each participant.

11. Results

11.1. Average performance and exclusions

As would be expected based on models of the ANS, participants performed quite accurately. Before outlier removal, there was an overall correlation of $r = .819$ between the true number of objects depicted and participants' responses to the images. To manage outliers, we again used the MAD (Leys et al., 2013). We determined the median average error rate across all participants separately for each image set (children's books: 5.0%, photographs: 13.1%), then removed participants whose average error was more than three times the MAD (children's books: 3.1%, photographs: 2.4%) away from the median error for that condition. This resulted in the removal of 4 participants from the counting book condition and 5 participants from the photographs condition, leaving N = 24 in the former and N = 23 in the latter for analyses (see Fig. 7, top row for summary models over trimmed responses). Note, however, that the exclusion of these outliers did not affect the significance of any reported statistical results in any of the behavioral experiments.

As can be seen with the trendlines in Fig. 7, participants tended to slightly underestimate (Izard & Dehaene, 2008). Participants showed a small amount of error even on the smallest values (1−3), but more importantly, they showed predictable increases in the variability of their responses as the numbers to be estimated got larger. This linearly increasing error—scalar variability—is a behavioral signature of the ANS that is *not* found with the exact representations used in the Object Tracking System (Feigenson & Dehaene, & Spelke, 2004). Therefore, our behavioral results substantiate our claim that the ANS is being activated during estimation of

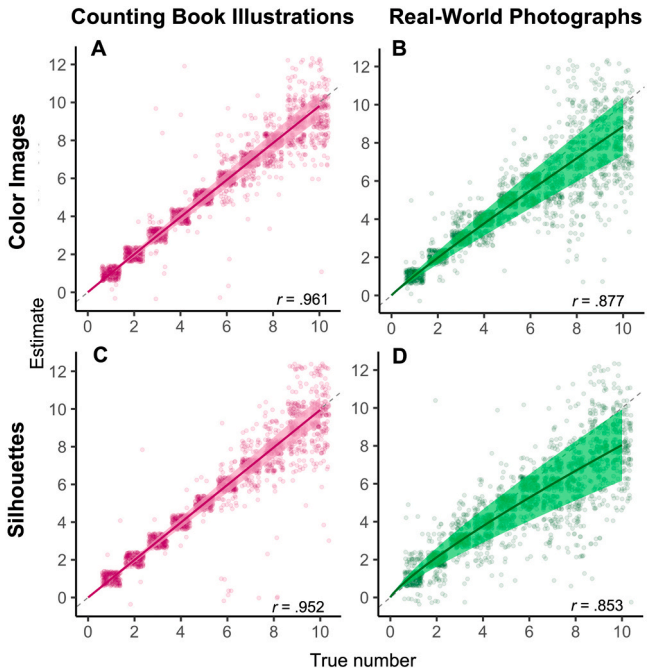

Fig. 7. Trimmed responses to stimuli in Experiments 3 (color images, top row) and 4 (silhouettes, bottom row). *Note.* Points are jittered on both axes; Lines represent average PsiMLE model fit per condition, with ribbon representing one σ.

these images.

11.2. Impact of non-numerical features on number estimates

We investigated the extent to which non-numerical features influenced number estimates, separately for each stimulus set. That is, can we predict what number participants will provide for their estimates given the continuous features of the ensemble? And are non-numerical features a better predictor of number estimate than number itself? We tested this question using linear regression and confirmed the results with partial correlations. However, because we did not directly manipulate the continuous features in our stimuli to test their impact on behavior, any relationship we find between the non-numerical features and our participants' responses should be taken as correlational rather than causal evidence. As in Experiments 1 and 2, due to the high collinearity between the non-numerical features, we used the non-numerical feature PCs as predictors to investigate how much variance in estimates is accounted for by the group of non-numerical features as a whole. We used PCs built from the group-level features, as that was the best performing group in the previous experiments.

In the counting book illustrations condition, a model predicting number estimates from the three (group-level feature) PCs found that all three were predictive (PC1: $B = -.33$, $SE\ B = .04$; PC2: $B = -1.62$, $SE\ B = .09$; PC3: $B = .77$, $SE\ B = .17$; $R^2 = .173$). Notably, the model only explained 17.3% of the variance in responses. We then compared this model to one using only the true number of objects in the image as a predictor; once again we used AIC and BIC to perform this comparison. The number model, $B = 0.95$, $SE\ B = 0.01$, $R^2 = .924$, AIC = −926.74, BIC = 4540.66, strongly outperformed the full non-numerical feature model (AIC = 3649.10, BIC = 9127.62). This result suggests that our participants' number estimates are more parsimoniously explained as being related to the actual number of objects on the page, and not the non-numerical features of those objects.

To see whether these features accounted for variance in estimates even when the true number was taken into account, we also performed partial correlations using the three PCs and number. We note, however, that in this case there is not much remaining variance to account for once the true number is taken into account. The partial correlation for number taking into account the three PCs was extremely large, $\rho = .953$, $p < .001$. Unsurprisingly, none of the partial correlations for any of the three PCs was significant, $ps > .169$.

For the photographs condition, a model predicting number estimates from the three PCs found that all three were predictive (PC1: $B = .34$, $SE\ B = .03$; PC2: $B = 3.69$, $SE\ B = .12$; PC3: $B = 1.43$, $SE\ B = .17$; $R^2 = .329$). We again compared this full model to a model that includes only the true number of objects in the image as a predictor. The model predicting number estimated from true number provided a good fit to the data, $B = 0.75$, $SE\ B = 0.01$, $R^2 = .667$, AIC = 1885.70, BIC = 8261.7, and again significantly outperformed the non-numerical feature model (AIC = 3459.49, BIC = 9846.9).

As in the counting book condition, we performed partial correlations to see whether these features accounted for variance in estimates even when the true number was taken into account. The partial correlation for number taking into account the three PCs was the largest, $\rho = .718$, $p < .001$, although with this stimulus set, the correlations for all three PCs (PC1: $\rho = -0.09$; PC2: $\rho = 0.09$; PC3: $\rho = .058$) were significant as well, $ps < .006$.

Although the non-numerical features in these images only accounted for a relatively small amount of variance in participants' estimates—between 17% and 33%—this amount was nonetheless significant, as were the partial correlations in the photograph

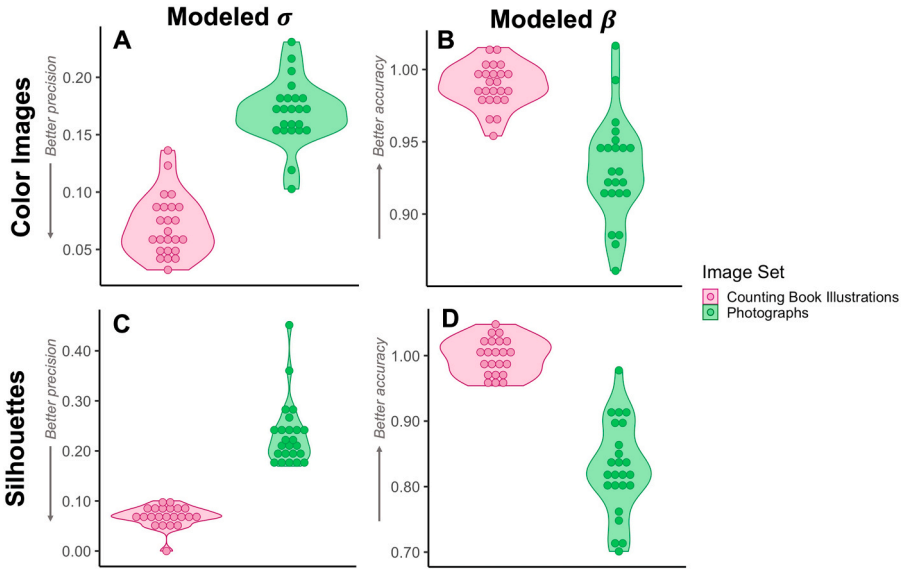

Fig. 8. Comparing model fit parameters from Experiment 3 (color images, top row) and Experiment 4 (silhouettes, bottom row) of counting book illustrations (pink) and photographs (green).

condition. Therefore, we find some (non-causal) evidence that non-numerical features may influence number estimates, albeit with a limited predictive ability compared to the predictive power of the true number of objects.

11.3. PsiMLE model fits

Next, we used PsiMLE to fit each participant's responses individually, resulting in a separate value of β and σ for each participant. If non-numerical features are necessary for number perception, we would expect to find superior performance in the photographs image set, where non-numerical features are more predictive of number. Strikingly, we found the opposite pattern for both model-fitted parameters. Participants who estimated based on counting book illustrations were significantly more precise (smaller σ: M =.070, SD =.026) than those who estimated based on photographs (M =.170, SD =.028), $t(44.52) = 12.62$, $p < .001$. They were also more accurate (β closer to 1) in the counting book condition (M =.989, SD =.015) compared to the photograph condition (M =.931, SD =.035), $t(29.35) = 7.34$, $p < .001$, (see Fig. 8, top row). This result strongly contrasts with the prediction of a non-numerical feature-based extraction algorithm for number, where performance should have been *better* for photographs than counting book illustrations.

12. Discussion

In this experiment, we found that human adults are able to perceive and report the number of objects in visually-complex natural scenes. In fact, they did so with similar (or even superior) precision and accuracy to that found in previous research with much more simple and controlled dot stimuli (Odic et al., 2016).

If non-numerical features were the basis of number extraction, performance would have been better in the photograph condition than the counting books condition. This was not the case. In fact, we found consistent evidence of the opposite effect: participants were more accurate to estimate the number of objects in counting book illustrations, where the relationships between number and non-numerical features were *weaker* than in the photographs.

One low-level confound that could potentially explain this pattern of results is that the background was generally more complex in the photographs compared to the counting books. This could lead to differences in the ease with which the target objects could be selected from the background, which may result in impaired extraction of the non-numerical features. If number estimates are built from non-numerical features, this could result in impaired number estimation in the photograph condition compared to the counting book condition (in which backgrounds are generally less complex). We control for this difference in Experiment 4.

13. Experiment 4

One possibility is that the pattern we found in Experiment 3 is due to confounding factors unrelated to the difference in the relationship between non-numerical features and number. One obvious confound is that photographs tend to be much more visually complex than the counting book illustrations. This could lead to a more difficult extraction process, which could artificially depress estimation performance in the photograph condition compared to the counting book condition. In Experiment 4, we tested this with a control experiment where the background of the images was equated in complexity: that is, we removed the backgrounds entirely, leaving only black silhouettes (which we had used to extract continuous features in Experiments 1 and 2). With this manipulation, non-numerical features are maintained and extremely easy to extract, and therefore we would expect to see better performance in the photographs condition than the counting books condition if those features are being used to make estimates.

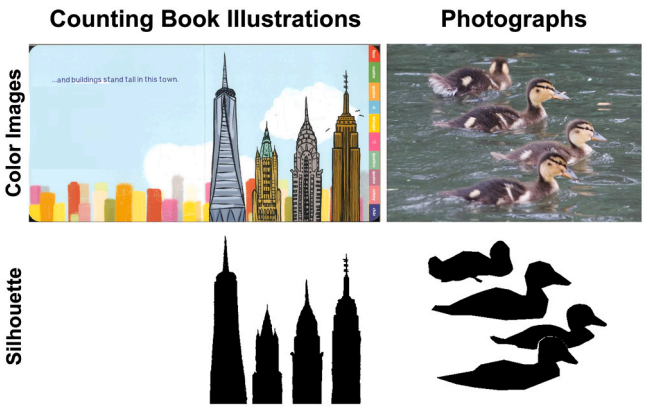

Fig. 9. Silhouettes and their original color images for stimuli in Experiment 4.

14. Method

14.1. Participants

A total of 51 people participated in this experiment (26 in the counting book condition and 25 in the photograph condition). Participants were recruited from Prolific and were paid at a rate of approximately $14 per hour for their participation. Demographic information about participants was not collected.

14.2. Materials

We used the silhouettes of the exact same stimuli that were used in Experiment 3. There were once again 80 images in each condition. In these silhouette images, the target objects are composed of black pixels and the entire background is made of white pixels (see Fig. 9). This manipulation removes all background details; if we still find a difference in precision between counting book and photograph estimates with these stimuli, it cannot be explained by a difference in background complexity.

14.3. Procedure

The procedure was identical to Experiment 3, except that participants were told before the experiment that they were going to be making estimates based on silhouettes rather than full color images.

15. Results

15.1. Exclusions

Overall, participants once again did quite well, with those in the counting book condition showing a correlation of $r = .914$ between the true number of objects in each stimulus and responses to each image, and those in the photograph showing a correlation of $r = .848$ (see Fig. 7, bottom row). Across conditions, participants again displayed scalar variability, consistent with relying on the ANS, and had a median error of 13.6% (approximately an error of responding 6 when the correct response was 7). As in the previous experiments, we removed any participants whose error rate was further than three times the MAD away from the median error rate for that condition. This exclusion removed 4 participants from the counting book condition and 1 from the photograph condition. Again, the removal of these participants did not affect the significance of any statistical results.

15.2. Model fits

We again fit each participant with a β and σ using PsiMLE's power function (Odic et al., 2016)—which measures the scalar variability and bias consistent with the ANS. Consistent with the results of Experiment 3, we found that performance was better in the counting book condition. Participants were more precise (lower σ) in the counting books condition ($M = .07$, $SD = .02$) compared to the photograph condition ($M = .23$, $SD = .06$), $t(28.43) = 12.10$, $p < .001$. They were also more accurate (β closer to 1) in the counting book condition ($M = 1.00$, $SD = .03$) compared to the photograph condition ($M = .83$, $SD = .07$), $t(30.53) = 11.13$, $p < .001$ (see Fig. 8, bottom row).

16. Discussion

This control experiment was intended to eliminate the complex backgrounds of the photographs while retaining or even highlighting the non-numerical features of the target objects. Therefore, if the superior performance from counting book illustrations was simply due to the ease with which the ensembles could be rapidly selected from the background, we would have expected this advantage to disappear in the silhouette condition. Instead, we found that the difference was maintained: estimation was more precise and more accurate based on counting book illustrations as compared to photographs. Again, this is the opposite of what would be expected if participants are using non-numerical features to derive their number estimates. Therefore, this result provides further support against that hypothesis. Human adults do not appear to rely on non-numerical features to estimate the number of objects in an image.

A natural next question to ask is whether adults, through extensive experience, have learned to modify their numerical extraction strategies away from a native bias to use non-numerical features. Therefore, we next asked whether young children would show similar patterns of results in their numerical estimation.

17. Experiment 5

Here, we ask whether children are more prone to using non-numerical features to estimate the number of objects in a visual scene than adults were. If they are, we would expect them to perform better in the photographs condition (where non-numerical features pattern more strongly with number) compared to the counting book illustrations condition. If we were to find this result, it might indicate that the pattern of performance we found in adults is due to an experiential modification of the number extraction algorithm,

which is initially predisposed to rely on non-numerical features. If not, it would suggest that children, like adults, do not rely on non-numerical features to estimate number.

18. Method

18.1. Participants

Participants were 94 children (M_{age} = 7;2 years, SD_{age} = 1;5, range 5;1–9;10) who were part of the database for the Laboratory for Child Development at Johns Hopkins. Most children were recruited for previous studies, through online advertisements and in-person recruitment at fairs and Farmer's Markets. The study was approved by the Homewood Institutional Review Board at Johns Hopkins University. Prior to participation, informed consent was obtained from a parent or guardian. Participants were compensated with a $5 Amazon gift card for their participation.

18.2. Materials

The materials were identical to those used in Experiments 3 and 4. There were 80 images from each image set (8 instances each of 1–10). There were two versions of each image (full color and silhouette); as in the adult experiments, each participant only saw images from one image set (color photographs, color counting books, silhouette photographs, and silhouette counting books).

18.3. Procedure

Overall, the experimental procedure was very similar to that used with adults, with a few modifications to make it accessible to children. Rather than running the experiment over the internet on the participant's own computers, an experimenter loaded the experiment on their own laptop and shared their screen with the child through a video call. This allowed the experimenter to control the progression of the experiment while still allowing the child to see the images.

At the beginning of each trial, the experimenter read the word indicating what would appear in the next image, and did not advance to the stimulus display until the child was ready and looking at the screen. Each stimulus image was displayed for 800 ms. Following the stimulus display, a slider appeared on the screen. The slider ranged from 0–20, as pilot testing indicated that children sometimes responded values larger than 12, and the initial position was set to 0. The experimenter waited until the child responded, prompting them to give an estimate if they were slow to respond (e.g., "How many BUSES do you think there were?"). If the child responded hesitantly, the experimenter asked them to confirm their answer before proceeding. Once the child confirmed their answer, the experimenter moved the slider to the response value and submitted the response, starting the next trial. Children were given the opportunity to take a break halfway through the experiment (after the completion of 40 trials). Each session was video recorded.

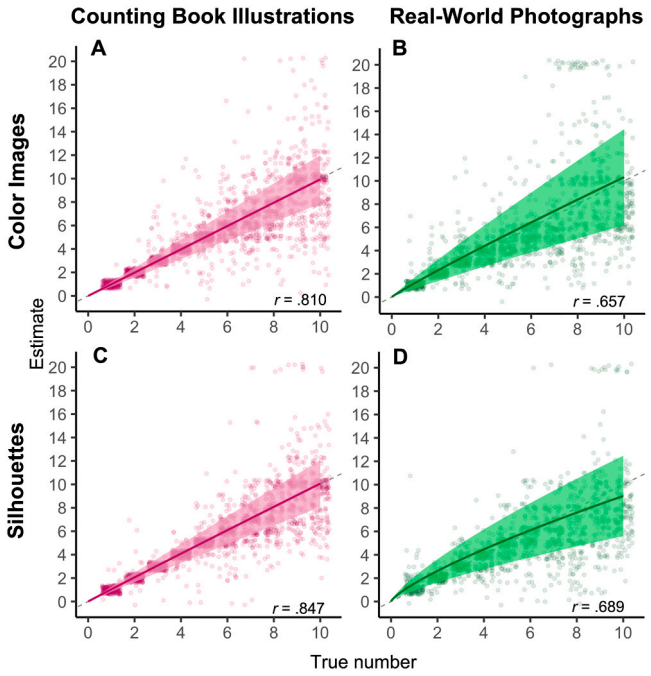

Fig. 10. Trimmed responses from all participants in each estimation condition (left: counting books, right: photographs). *Note.* Points are jittered on both axes; Lines represent average PsiMLE model fit per condition, with ribbon representing one σ.

18.4. Modeling

Again, we used the power model of PsiMLE to fit each child's responses, resulting in an estimate of their β and σ (Odic et al., 2016).

19. Results

19.1. Exclusions

Overall, children unsurprisingly performed less accurately than adults (see Fig. 10). Children, like adults, displayed the scalar variability that is the hallmark of relying on the ANS, and their median error across all trials was 21.2%, or an approximate error of responding 6 when the correct answer was 5. As in the previous experiments, we removed any participants whose error rate was further than three times the MAD away from the median error rate for that condition. This exclusion removed 13 participants who ranged from 37–100% average error (1 from the counting books color condition, 5 from the photographs color condition, 2 from the counting books silhouette condition, and 5 from the photographs silhouette condition). This left N = 23 in the counting books color condition, N = 18 in the photographs color condition, N = 23 in the counting books silhouette condition, and N = 17 in the photographs silhouette condition. Again, the removal of these participants did not affect the significance of any of the statistical results.

19.2. PsiMLE model fits

Once again, we fit each participant's responses with the power model of PsiMLE. We used two-way ANOVAs (image set X color condition) to predict model-fitted σ and β. For σ, consistent with the adult results, we found a main effect of image set, $F(1,77) = 59.34$, $p < .001$, where precision was better for counting books ($M = .18$, $SD = .09$) than for photographs ($M = .33$, $SD = .08$; see Fig. 11 left). There were no main effects or interactions with color condition, $ps > .189$.

For model-fitted β, there was again a main effect of image set, $F(1,77) = 16.73$, $p < .001$, where β was significantly larger (closer to 1) for those in the counting books condition ($M = 1.00$ $SD = .09$) than those in the photograph condition ($M = .89$, $SD = .14$; see Fig. 11, right). The main effect of color condition was not significant, $p = .112$. Finally, there was a weakly significant interaction between image set and color condition, $F(1,77) = 4.45$, $p = .038$. There was little change in β for counting books from the color ($M = .99$, $SD = .09$) to the silhouette condition ($M = 1.00$, $SD = .10$), while there was a large decrease in β for photographs from color ($M = .94$, $SD = .15$) to silhouettes ($M = .84$, $SD = .11$). We speculate that non-numerical features may be easier to extract in the silhouettes condition compared to the color photographs condition, in which case this is consistent with the suggestion that non-numerical features are a *distraction* rather than an aid to numerical processing in these images.

19.3. Discussion

We found that children showed highly similar response patterns to those of the adults in Experiments 3 and 4. In particular, children were both more precise and more accurate when estimating the number of objects in the illustrations of counting books compared to photographs. Again, this pattern of results is the *opposite* of what would be expected if children were using non-numerical features to estimate the number of objects in visual scenes. Taken together with the behavioral results from adult participants, these results strongly indicate that non-numerical features do not serve as the basis of numerical estimation, even for young children with relatively little estimation experience.

20. General discussion

In this series of experiments, we have demonstrated that counting book illustrations deviate from the statistics of the natural world in their depictions of visual ensembles. We raise the intriguing possibility that such pedagogical materials are *intentionally* designed to

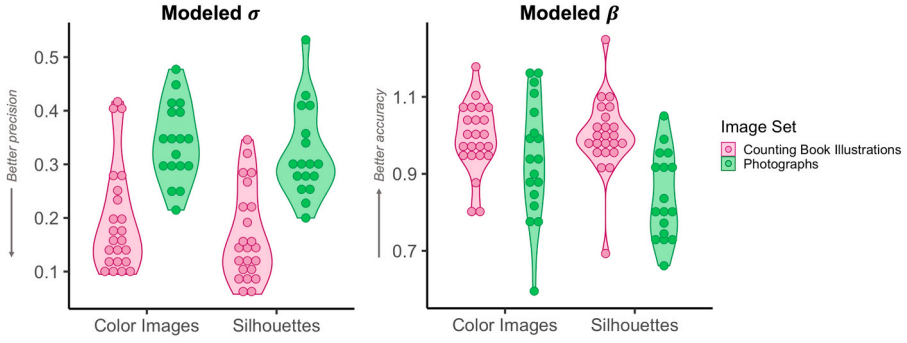

Fig. 11. Consistent with the results from adult participants, children performed better in the counting book condition (pink) than in the photograph condition (green), for both color images and silhouettes.

suppress such relationships. Given that we found superior number estimation performance from counting books as compared to photographs, this lack of relationship may be an intentional choice to *facilitate* number estimation. This would make counting books an example of a pedagogical material that has evolved to better suit human needs.

Although it had not been empirically validated, the claim that non-numerical features pattern with number in the real world is ubiquitous in the number literature (e.g., Abalo-Rodríguez et al., 2022; Leibovich et al., 2017; Smets et al., 2015; van Rinsveld et al., 2020). Here, we found that group-level features such as convex hull, surface area, and perimeter, which are often implicated in the number literature as driving numerical responses during discrimination tasks (Clayton et al., 2015; Gebuis et al., 2016), are more predictive than average object-level features or density in photographs of real-world scenes. Interestingly, we found that non-numerical features are far less predictive of number across the pages of children's books. This result would be surprising if non-numerical features were necessary or helpful for number perception, since the purpose of counting book illustrations is to provide a visual depiction of number: one would expect that whatever features are necessary for visual number perception would be highlighted in these materials.

Does this mean that counting book illustrations are poorly designed to allow for number perception? Our behavioral experiments indicate the opposite. We found that both adults and children are better (both more precise and more accurate) when estimating the number of objects in counting book illustrations than in photographs, and this was even true when accounting for differences in background complexity between the image sets. A possibility raised by these results is that counting books illustrations may in fact be *particularly well designed* for number learning about visual number. For instance, perhaps the lack of relationship between number and non-numerical features across the pages of counting books allows children to home in on the target dimension—number—without having to differentiate it from consistently confounding non-target features such as convex hull.

In both datasets, we found that density was less successful at predicting number than the other combination models, which is surprising given its prevalence in the number perception literature (Dakin et al., 2011; Durgin, 1995). However, given that density is defined using very different methods across the field (such as spatial frequency ratios, most commonly used with much smaller objects than those seen here), this claim need only apply to the particular definition of density that we employed (e.g., DeWind et al., 2015).

Another difference between the counting book illustrations and the real-world photographs is that the photographs tended to have more overlapping objects than did the illustrations (37% versus 16% of images). By visual inspection, the gaps between overlapping objects in the real-world photographs appear smaller and more variable on average than those we inserted into the images in the counting book illustrations. That these object outlines were designed using different methods in the two experiments is a limitation of our approach; although we do not believe it strongly impacts our conclusions, future work investigating this question should strive to use the same conversion method for any image sets being compared.

The difference between image sets in the percent of images containing overlapping objects is likely a source—if not *the* source—for the differences we found in behavioral performance between these two datasets. This is consistent with prior work demonstrating that people tend to underestimate more when objects are clustered together in a stimulus (e.g., Im et al., 2016). However, this difference may be consistent with our broader claims of the reduced relevance of non-numerical dimensions for extracting number. Why? Interestingly, it is likely that some or all of the non-numerical magnitudes we discussed, especially surface area, are *easier* for people to extract from images where the objects overlap considerably (as is often the case in area perception tasks, e.g., Odic et al., 2013). So even if the difference in overlap is the main source of difference in behavior, it's consistent with our claim that continuous features—which are easy, or even easier, to extract in the photographs—are *not* the main mechanism through which people are generating their number estimates.

Our analysis focused only on the retinotopic size of objects—that is, how much space the object occupies on the retina from a standard viewing distance. It is notable that this does not have to correspond to the "real-world" size of the objects depicted, nor their representation in brain regions that take into account real-world size, such as in the occipital cortex (Konkle & Oliva, 2012). Because the ANS is supposed to work incredibly rapidly, and because the literature has primarily focused on retinotopic features rather than real-world object sizes (e.g., Clayton et al., 2015; Gebuis et al., 2016; Smets et al., 2015; Szűcs et al., 2013), we chose to focus on those features for this analysis. Future research could investigate to what extent features such as real-world size influence numerical estimation, and whether it might drive behavior differently from retinotopic image size.

In the behavioral experiments in this paper, using the numerosities of 1 through 10, we observe scalar variability in our participants' responses. That is, they show increasing variability in their responses as the number to be estimated gets larger, which is the hallmark of relying on the ANS (Dehaene, 1997; Feigenson et al., 2004; Halberda et al., 2008; Piazza, 2010; Piazza et al., 2004). This demonstrates the relevance of our results to theorizing about the ANS. Nonetheless, the ANS extends to much higher values (up to at least 64, depending on density; Anobile et al., 2014). Therefore, future investigations of the relationship between number and continuous features should look at larger quantities than just 1–10, as it is an open question whether the conclusions we found here would hold for even larger quantities.

A potential limitation of this work is the size of our samples (between $N = 17$ and $N = 24$ for a given behavioral experiment after exclusions), as well as the between-subjects experimental design. Future work would improve the robustness of our conclusions by increasing the sample size and employing a within-subjects design (for both image sets and color conditions).

Our behavioral results are consistent with some previous evidence that number estimation is less impacted by non-numerical features than is number discrimination (Smets et al., 2015). What could be the locus of interference such that discrimination is impacted but estimation is not? Perhaps response conflict at the decision stage of a discrimination task drives results that differ by non-numerical congruity. Our results strongly suggest that the content of the representation itself is numerical and NOT an amalgamation of information from multiple non-numerical features.

In conclusion, the general claim that non-numerical features pattern with number in the real world is at least partially substantiated

by our data. However, these features pattern much less with number in pedagogical materials designed to facilitate number learning, children's counting books. We found that approximate perception of numerosity in counting book illustrations outperforms numerosity perception in real-world photographs. Therefore, we claim that even a combination of non-numerical ensemble features cannot serve as the perceptual basis for number perception.

Conflicts of Interest

The authors have no conflicts of interest to disclose.

Funding

Data collection and analysis were supported by an NSF GRFP, DGE1746891 awarded to E.M.S., and a McDonnell Foundation Scholar Award awarded to J.H.

CRediT authorship contribution statement

Sanford Emily Mae: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Halberda Justin:** Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Data Availability

Linked to GitHub.
Counting book illustrations (color and silhouette), feature extraction code (Original data) (Github)

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cogdev.2023.101415.

References

Abalo-Rodríguez, I., De Marco, D., & Cutini, S. (2022). An undeniable interplay: Both numerosity and visual features affect estimation of non-symbolic stimuli. *Cognition, 222*(February 2019), Article 104944. https://doi.org/10.1016/j.cognition.2021.104944

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Anobile, G., Cicchini, G. M., & Burr, D. C. (2014). Separate mechanisms for perception of numerosity and density. *Psychological Science, 25*(1), 265–270. https://doi.org/10.1177/0956797613501520

Ariely, D. (2001). Seeing sets: representation by statistical properties. *Psychological Science, 12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Butterworth, B. (2010). Foundational numerical capacities and the origins of dyscalculia. *Trends in Cognitive Sciences, 14*(12), 534–541. https://doi.org/10.1016/j.tics.2010.09.007

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature Human Behaviour, 4*(12), 1265–1272. https://doi.org/10.1038/s41562-020-00946-0

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Clayton, S., & Gilmore, C. (2015). Inhibition in dot comparison tasks. *ZDM - Mathematics Education, 47*(5), 759–770. https://doi.org/10.1007/s11858-014-0655-2

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

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin and Review, 8*(4), 698–707. https://doi.org/10.3758/BF03196206

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences of the United States of America, 108*(49), 19552–19557. https://doi.org/10.1073/pnas.1113195108

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press.

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

Durgin, F. H. (1995). Texture density adaptation and the perceived numerosity and distribution of texture. *In Journal of Experimental Psychology: Human Perception and Performance* (Vol. 21,(Issue 1), 149–169. https://doi.org/10.1037/0096-1523.21.1.149

Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology, 18*(18), 855–856. https://doi.org/10.1016/j.cub.2008.07.053

Elia, I., van Den Heuvel-Panhuizen, M., & Georgiou, A. (2010). The role of pictures in picture books on children's cognitive engagement with mathematics. *European Early Childhood Education Research Journal, 18*(3), 275–297. https://doi.org/10.1080/1350293X.2010.500054

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants' manual search. *Developmental Science, 6*(5), 568–584. https://doi.org/10.1111/1467-7687.00313

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: object files versus analog magnitudes. *Psychological Science, 13*(2), 150–156. https://doi.org/10.1111/1467-9280.00427

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

Feigenson, L., & Halberda, J. (2004). Infants chunk object arrays into sets of individuals. *Cognition, 91*(2), 173–190. https://doi.org/10.1016/j.cognition.2003.09.003

Ganea, P., Pickard, M. B., & DeLoache, J. (2008). Transfer between picture books and the real world by very young children. *Journal of Cognition and Development, 9*(1), 46–66. https://doi.org/10.1080/15248370701836592

Gebuis, T., Cohen Kadosh, R., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: a critical review. *Acta Psychologica, 171*, 17–35. https://doi.org/10.1016/j.actpsy.2016.09.003

Göbel, S. M., McCrink, K., Fischer, M. H., & Shaki, S. (2018). Observation of directional storybook reading influences young children's counting direction. *Journal of Experimental Child Psychology, 166*, 49–66. https://doi.org/10.1016/j.jecp.2017.08.001

Goldstein, A., Cole, T., & Cordes, S. (2016). How parents read counting books and non-numerical books to their preverbal infants: an observational study. *Frontiers in Psychology, 7*(JUL), 1–10. https://doi.org/10.3389/fpsyg.2016.01100

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe, & L. Robertson (Eds.), *From perception to consciousness: Searching with Anne Treisman* (pp. 1–21). Oxford University Press.

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. 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

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science, 17*(7), 572–576. https://doi.org/10.1111/j.1467-9280.2006.01746.x

Im, H. Y., Zhong, S. hua, & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision Research, 126*, 291–307. https://doi.org/10.1016/j.visres.2015.08.013

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

Jones, E., Oliphant, T., Peterson, P. , 2001. SciPy: Open Source Scientific Tools for Python.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron, 74*(6), 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: an investigation of the conceptual sources of the verbal counting principles. *Cognition, 105*(2), 395–438. https://doi.org/10.1016/j.cognition.2006.10.005

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": the role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences, 40*. https://doi.org/10.1017/S0140525×16000960

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. 2014. Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8693 LNCS(PART 5), 740–755. https://doi.org/10.1007/978–3-319–10602-1_48.

Liu, R. X., Kuang, J., Gong, Q., & Hou, X. L. (2003). Principal component regression analysis with SPSS. *Computer Methods and Programs in Biomedicine, 71*(2), 141–147. https://doi.org/10.1016/S0169-2607(02)00058-5

Morgan, M. J., Raphael, S., Tibber, M. S., & Dakin, S. C. (2014). A texture-processing model of the "visual sense of number. *Proceedings of the Royal Society B: Biological Sciences, 281*(1790), 1–9. https://doi.org/10.1098/rspb.2014.1137

Næs, T., & Mevik, B. H. (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics, 15*(4), 413–426. https://doi.org/10.1002/cem.676

Odic, D., Im, H. Y., Eisinger, R., Ly, R., & Halberda, J. (2016). PsiMLE: a maximum-likelihood estimation approach to estimating psychophysical scaling and variability more reliably, efficiently, and flexibly. *Behavior Research Methods, 48*(2), 445–462. https://doi.org/10.3758/s13428-015-0600-5

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), 1103–1112. https://doi.org/10.1037/a0029472

Odic, D., & Oppenheimer, D. M. (2023). Visual numerosity perception shows no advantage in real-world scenes compared to artificial displays. *Cognition, 230*(April 2022), Article 105291. https://doi.org/10.1016/j.cognition.2022.105291

Piantadosi, S. T. (2016). A rational analysis of the approximate number system. *Psychonomic Bulletin and Review, 23*(3), 877–886. https://doi.org/10.3758/s13423-015-0963-8

Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences, 14*(12), 542–551. https://doi.org/10.1016/j.tics.2010.09.008

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

R Core Team. , 2013). R: A language and environment for statistical computing. Foundation for Statistical Computing. http://www.r-project.org/.

Rajan, P., Ma, Y., & Jedynak, B. (2019). Cox processes for counting by detection. *Journal of Mathematical Imaging and Vision, 61*(3), 380–393. https://doi.org/10.1007/s10851-018-0838-5

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: what children must learn and when they learn it. *Cognition, 108*(3), 662–674. https://doi.org/10.1016/j.cognition.2008.05.007

Smets, K., Sasanguie, D., Szűcs, D., & Reynvoet, B. (2015). The effect of different methods to construct non-symbolic stimuli in numerosity estimation and comparison. *Journal of Cognitive Psychology, 27*(3), 310–325. https://doi.org/10.1080/20445911.2014.996568

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science, 133*(3446), 80–86. https://doi.org/10.1126/science.133.3446.80

Sun, Z., & Firestone, C. (2021). Curious objects: how visual complexity guides attention and engagement. *Cognitive Science, 45*(4). https://doi.org/10.1111/cogs.12933

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4-5-year-old children. *Developmental Science, 18*(4), 556–568. https://doi.org/10.1111/desc.12239

Szűcs, D., Nobes, A., Devine, A., Gabriel, F. C., & Gebuis, T. (2013). Visual stimulus parameters seriously compromise the measurement of approximate number system acuity and comparative effects between adults and children. *Frontiers in Psychology, 4*(JUL), 1–12. https://doi.org/10.3389/fpsyg.2013.00444

Testolin, A., Zou, W. Y., & McClelland, J. L. (2020). Numerosity discrimination in deep neural networks: initial competence, developmental refinement and experience statistics. *Developmental Science, 23*(5), 1–13. https://doi.org/10.1111/desc.12940

van den Heuvel-Panhuizen, M., van de Boogaard, S., & Doig, B. (2009). Picture books stimulate the learning of mathematics. *Australian Journal of Early Childhood, 34*(3), 30–39. https://doi.org/10.1177/183693910903400305

van Rinsveld, A., Guillaume, M., Kohler, P. J., Schiltz, C., Gevers, W., & Content, A. (2020). The neural signature of numerosity by separating numerical and continuous magnitude extraction in visual cortex with frequency-tagged EEG. *Proceedings of the National Academy of Sciences of the United States of America, 117*(11), 5726–5732. https://doi.org/10.1073/pnas.1917849117

vanMarle, K., Chu, F. W., Mou, Y., Seok, J. H., Rouder, J., & Geary, D. C. (2018). Attaching meaning to the number words: contributions of the object tracking and approximate number systems. *Developmental Science, 21*(1), 1–17. https://doi.org/10.1111/desc.12495

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243. https://doi.org/10.1037/a0027127

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: the role of statistical perception in determining whether awareness overflows access. *Cognition, 152*, 78–86. https://doi.org/10.1016/j.cognition.2016.01.010

Ward, J. M., Mazzocco, M. M., Bock, A. M., & Prokes, N. A. (2017). Are content and structural features of counting books aligned with research on numeracy development? *Early Childhood Research Quarterly, 39*, 47–63. https://doi.org/10.1016/j.ecresq.2016.10.002

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: the psychophysics of number representation. *Psychological Science, 10*(2), 130–137. http://www.jstor.org/stable/40063393.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74*(1), 1–11. https://doi.org/10.1016/S0010-0277(99)00066-9