



# A number-line task with a Bayesian active learning algorithm provides insights into the development of non-symbolic number estimation

Sang Ho Lee<sup>1</sup> · Dan Kim<sup>1</sup> · John E. Opfer<sup>1</sup> · Mark A. Pitt<sup>1</sup> · Jay I. Myung<sup>1</sup>

Accepted: 11 November 2021 / Published online: 16 December 2021  
© The Psychonomic Society, Inc. 2021

## Abstract

To characterize numerical representations, the number-line task asks participants to estimate the location of a given number on a line flanked with zero and an upper-bound number. An open question is whether estimates for symbolic numbers (e.g., Arabic numerals) and non-symbolic numbers (e.g., number of dots) rely on common processes with a common developmental pathway. To address this question, we explored whether well-established findings in symbolic number-line estimation generalize to non-symbolic number-line estimation. For exhaustive investigations without sacrificing data quality, we applied a novel Bayesian active learning algorithm, dubbed *Gaussian process active learning* (GPAL), that adaptively optimizes experimental designs. The results showed that the non-symbolic number estimation in participants of diverse ages (5–73 years old,  $n = 238$ ) exhibited three characteristic features of symbolic number estimation.

**Keywords** Numerical cognition · Cognitive development · Cognitive modeling · Gaussian process · Active learning · Hierarchical Bayesian modeling

## Introduction

The number-line task is one of the most commonly used tasks to study numerical cognition. In a conventional task, participants estimate the location of a symbolic number (e.g., Arabic numeral of 8) along a line flanked by two numbers (e.g., 0 and 50). Performance on number-line tasks with symbolic numbers has been viewed as depicting the approximate numerical representations that are also shared with non-symbolic numbers (e.g., number of dots; Dehaene et al., 2008; Siegler & Opfer, 2003). In this view, the meanings of symbolic numbers are learned by mapping to their non-symbolic referents (e.g., mapping numeral 8 to 8 dots), forming an association between symbolic and non-symbolic number representations (Dehaene, 2011). Therefore, performance on symbolic number-line tasks is thought to predict how non-symbolic numbers would be estimated on number lines.

Against the shared-representation account, however, some suggest independent representations for symbolic

and non-symbolic numbers (Carey, 2004; Carey & Barner, 2019; Lyons et al., 2012; Rips et al., 2008). According to this perspective, symbolic-number learning does not require mapping to the cardinality of a set from early development (Carey & Barner, 2019; Rips et al., 2008). Alternatively, symbolic numbers may be mapped to their non-symbolic referents early but become “estranged” from them late in development (Lyons et al., 2012). These accounts would not predict major findings in symbolic number-line studies to be observed in number-line tasks with non-symbolic numbers. Despite its theoretical importance, whether the characteristic features of symbolic number estimation are also evident in non-symbolic number estimation has not been fully investigated.

One of the characteristic features of symbolic number-line estimation is the “log-to-linear shift” (Siegler et al., 2009). This shift refers to the observation that younger and/or less educated populations produce logarithmic estimates in situations where older and/or more educated populations produce more linear estimates (Berteletti et al., 2010; Opfer & Siegler, 2007; Siegler & Booth, 2004; Siegler & Opfer, 2003; Thompson & Opfer, 2008). A prominent interpretation of the log-to-linear shift is that it reflects adaptive changes in numerical representations. The default representation, shared across human and non-human species, is

✉ Sang Ho Lee  
lee.7285@osu.edu

<sup>1</sup> Department of Psychology, The Ohio State University,  
212 Psychology Building, 1835 Neil Avenue, Columbus,  
OH 43210, USA

logarithmically scaled. Here, the difference between 1 and 10 is subjectively larger than the difference between 101 and 110. With age and schooling, however, the formal properties of the decimal system are gradually learned and can be applied to new situations where the difference between 1 and 10 is equal to the difference between 101 and 110 (Dehaene, 2011). In this account, non-symbolic estimation would also be expected to show log-to-linear shifts over development (Kim & Opfer, 2018; Yuan et al., 2020).

A second feature of symbolic number-line estimation is that log and linear representations coexist (Siegler & Opfer, 2003; Thompson & Opfer, 2010). The coexistence of different representations is typically explored by manipulating the number range in number-line tasks (e.g., estimating 20 on a 0–100 vs. 200 on a 0–1,000 number line). A common finding is that people are more likely to rely on a log representation for a larger number range (Opfer et al., 2019; Siegler & Booth, 2004; Siegler & Opfer, 2003). If the same representations were used for non-symbolic numbers, non-symbolic number estimates would also be more logarithmic when to-be-estimated numbers are large (e.g., 200 dots) rather than small (e.g., 20 dots). This question, however, has not been systematically explored.

A third feature of symbolic number-line estimation is that the linearity of estimates predicts proficiency with numbers in other contexts (Booth & Siegler, 2008; Siegler & Booth, 2004). The linearity of symbolic estimates predicts number memory (Opfer et al., 2019), counting (Östergren & Träff, 2013), math learning (Booth & Siegler, 2008; Siegler & Ramani, 2008), math scores (Fazio et al., 2014; Kim & Opfer, 2017), dyscalculia (Geary et al., 2008), and a genetic disorder, like Williams syndrome (Opfer & Martens, 2012). Whether the linearity of non-symbolic number estimates is associated with non-symbolic math is still an open question.

Despite previous attempts to address these questions using symbolic and non-symbolic number-line tasks, the findings in the literature are not consistent. For example, whereas symbolic and non-symbolic estimation appears to share the log-to-linear developmental trajectory in some studies (Kim & Opfer, 2018; Sasanguie et al., 2012; Sella et al., 2015), other studies report that non-symbolic estimation develops differently from symbolic estimation (Kolkman et al., 2013; Sasanguie et al., 2016). Table 1 summarizes previous research examining the development of symbolic and non-symbolic number estimation. As shown in the table, research parameters, including participants' ages, number scales, and applied models, were different across studies, which could have caused the discrepancy in the results. This state of affairs calls for a rigorous examination of the relations between the two types of number estimation.

The purpose of the present study, using a non-symbolic version of the number-line task, is to systematically investigate whether the preceding three features of symbolic

number estimation also characterize non-symbolic number estimation. Exploring these features in number-line tasks is inherently challenging because of the potential range dependency of number-line estimates. If the number range tested is too small or too large, developmental differences and the association between estimates and math skills may appear absent (Clarke et al., 2018). Determining appropriate number range can be more challenging in a non-symbolic number-line task. For symbolic number, there is a certain range of numbers that children more frequently encounter in a particular grade in education (e.g., one-digit numbers in kindergarten, two- and three-digit numbers in third grade). Symbolic number-line tasks with familiar number ranges might show individual variations that correlate with symbolic math competence. However, it is difficult to specify the range of non-symbolic numbers children come across in everyday life. Non-symbolic numbers vary greatly for every individual in size, from a relatively small number (e.g., number of cookies in a jar) to a very large number (e.g., number of people in a large stadium).

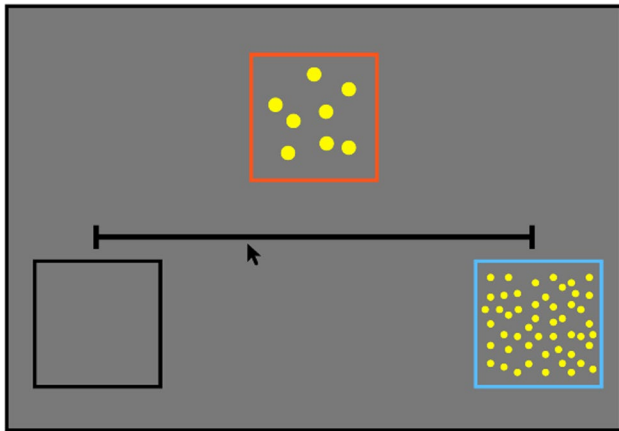
The number-line range is determined by the upper-bound number, which is often determined by an educated guess of investigators. When it is unclear which upper-bound number should be used, a simple but costly approach would be to examine estimates on multiple number-line tasks with different upper-bound numbers. The same approach has been used to explore the coexistence of log and linear representations, but with only a few different number ranges (e.g., 0–10 and 0–100; Berteletti et al., 2010). To fully examine the changes in estimates across number ranges, it would be desirable to use as many upper-bound numbers as possible. However, this approach would require many more trials than the typical task with a single, fixed scale (total trials = number of trials in each range  $\times$  number of upper bounds).

The present study aimed to address the design bottleneck in number-line tasks by applying a novel *multi-scale* non-symbolic number-line task (Fig. 1) that can be as brief as a fixed-scale number-line task but without sacrificing data quality. This highly efficient task is developed based on an algorithm that combines optimal experimental design in statistics (Atkinson & Donev, 1992) with active learning in machine learning (Settles, 2012). It provides a computational means to characterize non-symbolic number estimation over a much wider range of number scales for an individual, thereby quantifying developmental changes in estimates, as efficiently and accurately as possible. Specifically, in an experiment using a Bayesian active learning algorithm, dubbed *Gaussian process active learning* (GPAL), that our lab has developed (Chang et al., 2021), we delineate number-line estimation functions across four or more different number ranges, and compare the functions among individuals.

**Table 1** Summary of number-line studies with symbolic and non-symbolic number

Study	Participants	Number scale	Model applied	Major finding
Sasanguie et al. (2012)	Children in K-2nd, and 6th grades	0–10 for K–1st, 0–100 for 1st–2nd, 6th graders	Log, linear	Estimates were linear in the 0–10 task. For the 0–100 number lines, young children were more log than older children regardless of number formats
Kolkman et al. (2013)	Children repeatedly tested at ages 4, 5, and 6 years	1–100	Linear	Non-symbolic and symbolic estimation was loaded onto different factors in early development and integrated into one factor with age
Sella et al. (2015)	Children in K, 1st, 3rd grades	0–100	Log, linear	Estimates were more log in K and 1st graders than in 3rd graders for both symbolic and non-symbolic number
Honoré and Noël (2016)	Children in the last year of preschool	2–19	Results not reported	Non-symbolic number training did not affect estimation of symbolic number but improved that of large non-symbolic number. Symbolic number training improved estimation of symbolic number and small non-symbolic number
Maertens et al. (2016)	Children in kindergarten	0–10	No model applied	Number-line training with symbolic and non-symbolic number improved symbolic and non-symbolic estimation as well as arithmetic
Sasanguie et al. (2016)	Children in 1st, 2nd, and 6th grades	0–100	Log, linear, 2-phase linear, 0–2CPM	For all grades, symbolic number estimates were better fit by 0–2CPM. Non-symbolic estimates were log in 1st graders but better fit by 0CPM in older children
Kim and Opfer (2018)	Children aged 5–6 years and adults	0–30	MLLM, dynamic MLLM, Bayesian integration model	Non-symbolic estimates were more log than symbolic estimates, but both were better fit by MLLM or dynamic MLLM, showing log-to-linear changes with age
Yuan et al. (2020)	Children aged 4–6 years	0–1,000	Log, linear, MLLM, 1–2CPMs	CPMs did not provide good fits to estimates. Non-symbolic estimation was more log and correlated with symbolic estimation and counting when convex hull was controlled. Such associations between symbolic and non-symbolic estimation were not found when the convex hull was highly congruent with the non-symbolic number
Opfer et al. (2021)	Chinese and US children aged 5 years	0–30	MLLM, 1–2CPMs	Estimates of symbolic and non-symbolic numbers were better fit by MLLM and correlated with each other
van 't Noordende et al. (2021)	Children repeatedly tested at ages 3.5 and 5 years	0–100 for non-symbolic and 0–10 for symbolic number	No model applied	Non-symbolic number estimation at age 3.5 years predicted symbolic number estimation at age 5 years

K kindergarten, CPM cyclic power model, MLLM mixed log-linear model



**Fig. 1** Non-symbolic number-line task. *Note.* A given number (in the red box) is estimated on a line with zero and an upper-bound number (in the blue box)

We hypothesize that if non-symbolic numbers share representations with symbolic numbers, non-symbolic number-line estimation would exhibit the three features of symbolic number-line estimation described above: Non-symbolic estimates would change from log to linear with age, become more logarithmic in the same individuals as upper bounds increase, and show associations with non-symbolic math proficiency measured in a non-estimation context (approximate addition).

## Method

### Experiment

The present study applied GPAL to the non-symbolic version of the number-line task in Fig. 1 in order to infer psychophysical functions underlying non-symbolic number estimates in the two-dimensional design space. That is, the given number (design variable 1) and the upper-bound number (design variable 2; scale) were varied from one trial to another in an adaptive and optimal manner prescribed by the GPAL algorithm. This is unlike typical number-line tasks that manipulate only the given number during the task.

**Participants** Seventy-three children aged between 5 and 13 years (47 males,  $M_{age}$ : 8.69,  $SD_{age}$ : 2.09) were tested at a local science museum. 165 adults were recruited from a local university ( $n = 31$  (16 males),  $M_{age}$ : 19.44,  $SD_{age}$ : 0.99) and from Amazon Mechanical Turk (MTurk;  $n = 134$  (69 males),  $M_{age}$ : 39.25,  $SD_{age}$ : 12.91). All participants were recruited from the USA. Experiments in the current study were approved by the Institutional Review Board (IRB) in the local university.

**Bayesian power analysis** We determined the number of participants based on a Bayesian power analysis (Kruschke, 2010). The effect of interest was whether the logarithmicity component ( $\lambda$ ) in a mixed log-linear model (MLLM; described in a later section) changes with the upper-bound number across age groups. Statistical power was calculated as the estimated probability that new data replicate the difference in the values of  $\mu_\lambda$  (a hyperparameter that determines the prior mean of  $\lambda$ ) in a 0–50 number line ( $\mu_{\lambda,50}$ ) and a 0–100 number line ( $\mu_{\lambda,100}$ ), which was found in our pilot experiment (0.074 effect size). The effect was considered replicated if Bayesian 99% highest posterior density interval (HPDI) of  $\mu_{\lambda,100} - \mu_{\lambda,50}$  did not include zero. We set a desired level of statistical power at 0.8. Estimated power was 0.8 for 73 participants (children) and 0.88 for 165 participants (adults). When 95% HPDIs were used instead of 99% HPDIs, the power level was 1 for both children and adults.

**Stimuli and procedure** Each participant completed a non-symbolic number-line task (Fig. 1) and a non-symbolic math proficiency task – i.e., approximate addition. The number-line task was always given first. In this task, a group of dots was presented above a line for 2,000 ms every trial. Participants were asked to decide the location of the given number of dots on a line. The response was made by mouse-clicking the assumed location of the number on the line. The number of dots to estimate was chosen by GPAL every trial, among the integers between 5 and an upper bound. Small numbers (0–4) that are subitizable (Feigenson et al., 2004) were excluded from the design.

Besides given numbers, upper-bound numbers also varied trial by trial. The number of dots at the right end of a line (i.e., the upper bound) in Fig. 1 was chosen by GPAL every trial, whereas the number of dots at the left end (i.e., the lower bound) was always zero. The task for children had fewer total trials and fewer possible upper bounds than those for adults, to make the task suitable for children. The possible upper-bound values were 50, 100, 200, and 400 for children, and 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 for adults. The task consisted of 50 trials for children, and 90 trials for adults. The GP posterior was reset after the first half (block) of the trials, in order to test the reliability of GPAL by comparing the function estimates in the first block (i.e., test) and the second block (i.e., retest), each with a half of the trials (25 trials for children, 45 trials for adults). After each trial, participants had to press the spacebar to proceed to the next trial. There were five practice trials before the test block. We controlled for the size and the cumulative area of the dots in the given-number box. On half of the trials, the size of the dots was held the same in both the given-number and the upper-bound boxes, with the cumulative area increasing with the given number. On the other half, the cumulative area of the dots in the given-number

box was equal to that in the upper-bound box, with the dot size decreasing with the given number. Also, the dots of given- and upper-bound numbers were randomly spread in the boxes, such that the convex hull of the dots would not vary with numbers. According to Yuan et al. (2020), estimation of non-symbolic numbers is correlated with symbolic number estimation when perceptual cues, such as the convex hull, are not readily available. The random distribution of dots, therefore, was expected to elicit the use of representations and skills shared with symbolic numbers.

To test numerical performance outside estimation contexts, participants were also given a non-symbolic addition task upon completion of the number-line task. All adult participants completed both tasks. For children, we asked if they were willing to do another task when they completed the number-line task, to prevent loss of data quality due to fatigue. Fifty-one children chose to complete the approximate addition task. In this task, participants viewed two arrays of dots going into a box, consecutively for 1,500 ms per array, on the left side of the screen. Then, there was another array of dots presented on the right side of the screen for 1,500 ms. Participants were instructed to choose which side had more dots using the *f* or *j* key press. Each array had from 5 to 50 dots, and the sum of two arrays of dots in the gray box was less than 51. The array size was drawn from five ratio bins (1.05, 1.14, 1.2, 1.5, 2.0). There were 30 trials for children (six trials for each ratio bin), and 50 trials (ten trials for each ratio bin) for adults. There were two practice trials without feedback.

In what follows, we provide a brief overview of GPAL with which the experiment was conducted.

### Gaussian Process Active Learning (GPAL)

GPAL is an algorithm-based experimental method for adaptively optimizing experimental designs in order to infer an unknown function with the fewest possible number of observations (Chang et al., 2021). GPAL comprises the following three iterative steps that are performed on each trial: (1) *Design optimization* step in which the optimal design is identified based on the current state of knowledge about the model being inferred; (2) *Experiment* step in which stimuli are presented with the optimized design configuration and a response is observed; and (3) *Model inference* step in which the observed response is used to infer an updated model, which in turn becomes a new model for the next iteration.

GPAL has been developed as a nonparametric extension of a parametric Bayesian active learning algorithm for optimal experimental design, dubbed *Adaptive Design Optimization* (ADO; Cavagnaro et al., 2010). A key difference between GPAL and ADO is in the model inference step. ADO requires the assumption of a parameterized model that

specifies the functional form that generates responses in an experiment. For example, in the number-line task, the model might assume that number estimates follow log, linear, or cyclic power functions. In contrast, GPAL is “model-free” in that it does not make *a priori* assumptions about the possible model form, and instead, directly infers the form based on observed responses.

To infer the underlying functional form, GPAL utilizes a nonparametric Bayesian method known as a Gaussian process (GP) (e.g., Cox et al., 2012; Griffiths et al., 2009; Rasmussen & Williams, 2006; Schulz et al., 2018). Being nonparametric, GP is capable of modeling a virtually limitless range of functional forms without being subject to the constraints imposed by a parametric model family as in ADO. Formally, a GP is defined as a stochastic random process that forms a Gaussian distribution of functions over the function space (Rasmussen & Williams, 2006):

$$f(x) \sim GP(m(x), k(x, x')), \quad (1)$$

where  $f(x)$  is the underlying function to be inferred from observed data, and  $x$  and  $x'$  are two different points in the design space. In the above equation, the mean function  $m$  and the kernel function  $k$ , the latter of which governs the smoothness of the function  $f(x)$ , are defined as statistical expectations with respect to the distribution of functions:

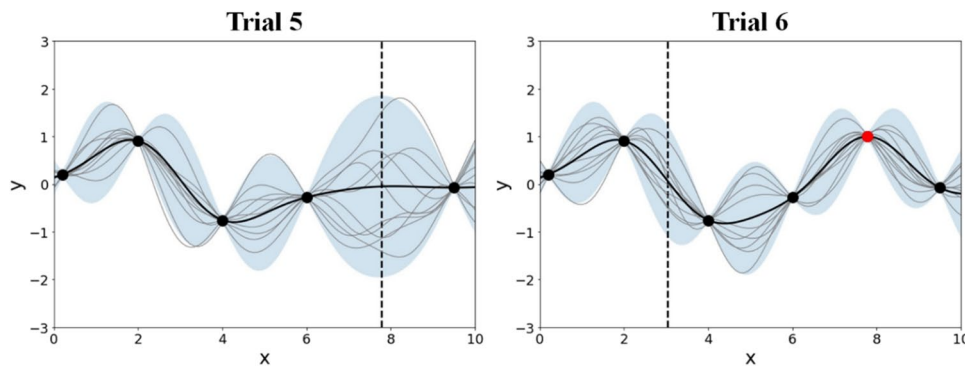
$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] \\ k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))). \end{aligned} \quad (2)$$

In the present study, we use the squared exponential kernel function.

Once the model is inferred by GP, it is then used in the design optimization step to identify the optimal design. This optimization process is known as *active learning* in machine learning (Settles, 2012). The active learning in GPAL relies upon the (Bayesian updated) distribution of the function  $f(x)$  in Eq. (1). In particular, we implement the uncertainty sampling scheme of Lewis and Catlett (1994) in which GPAL queries the design point  $x$  with the highest variance of  $f(x)$ . As such, an optimal design is the one that leads to the largest reduction in uncertainty about the unknown function  $f(x)$ . Shown in Fig. 2 is an example of the model inference and design optimization steps in GPAL.

### Models of numerical estimation

We used formal models of numerical estimation in literature to assess the characteristics of GPAL-inferred functions (i.e., GP-estimated posterior mean function). The three features of number-line estimates could be interpreted in a log-linear framework, but GP-estimated functions per se do not provide a quantifiable measure of logarithmicity. To quantify the degree of log compression in GPAL-inferred functions, we



**Fig. 2** An illustrated scheme of the Gaussian Process Active Learning (GPAL) algorithm. *Note.* The left panel summarizes the current state of knowledge about the underlying model after five observations. In this graph, the black dots are observed data, and the black curve is the inferred model as the GP mean function. The blue area indicates the two standard deviations from the GP mean and represents the uncertainty about the underlying model. The grey curves are a few

functions sampled from the GP. GPAL selects the optimal design for the next trial as the point with the largest variance, indicated by the dashed vertical black line. Shown on the right panel is a new observation (red dot) made with the optimized design, which is then used to update the GP. The iterative process repeats itself for the next trial. For additional technical details, readers are directed to Chang et al. (2021)

used a mixed log-linear model (MLLM; Anobile et al., 2012; Opfer et al., 2016; see below for details).

However, interpreting the results in a log-linear framework might not be justifiable if GPAL produces functions better explained by alternative models, such as cyclic power models (CPMs; Hollands & Dyre, 2000). The CPMs are derived from the proportional reasoning account, assuming log or linear patterns in estimates do not reflect numerical representations, but proportional reasoning skills required for number-line estimation (Barth & Paladino, 2011; Cohen & Sarnecka, 2014; Rouder & Geary, 2014; Slusser et al., 2013). Both the MLLM and CPMs can describe log and linear functions, but some versions of the CPM also predict distinctive patterns that the MLLM would not describe. Given that GPAL is free of assumptions about the shapes of underlying functions, whether logarithmic or cyclic, comparing the two models with the functions inferred by GPAL will allow us to determine the functional form of numerical estimation in an unbiased manner without favoring one model or another (see Appendix for GPAL simulations).<sup>1</sup>

**Mixed log-linear model (MLLM)** The MLLM is a commonly used model to describe number-line estimates. This model is defined as:

<sup>1</sup> Recent studies have suggested that the choice of given numbers in the number-line task could affect the comparison between the MLLM and the CPMs (Opfer et al., 2016; Slusser et al., 2013). For example, a task with given numbers evenly distributed across the number line (Slusser et al., 2013) is more likely to support the CPMs, whereas given numbers concentrated in the early part of the number line are in favor of the MLLM (Siegler & Opfer, 2003). However, the designs selected by GPAL are model-neutral, since GP infers the underlying psychophysical functions without a priori assumptions about their shape.

$$y = a \left( \lambda \frac{U}{\ln(U)} \ln(x) + (1 - \lambda)x \right) + b, \quad (3)$$

where  $y$  is the estimate of a given number  $x$  in a number line with the upper bound of  $U$ . The parameter  $\lambda$ , as a measure of logarithmicity, is the relative weight of the logarithmic function to the linear function in estimation. The estimate  $y$  becomes completely linear with  $\lambda = 0$ , and completely logarithmic with  $\lambda = 1$ .  $a$  and  $b$  are two scaling parameters.

The model parameters were estimated by fitting a hierarchical Bayesian model (Lee & Wagenmakers, 2014) using JAGS in MATLAB (Plummer, 2003; Steyvers, 2011). The MCMC sampling was iterated 150,000 times, with the first 50,000 samples being discarded as burn-in samples. The model parameters for participant  $i$  were defined as follows:  $\lambda_i \sim \text{Beta}(\mu_\lambda \eta_\lambda, (1 - \mu_\lambda) \eta_\lambda)$ ,  $a_i \sim \text{Beta}(1, 1)$ , and  $b_i \sim \text{Uniform}(0, U)$ , where  $\mu_\lambda$  and  $\eta_\lambda$  are hyperparameters following  $\mu_\lambda \sim \text{Beta}(1, 1)$ , and  $\eta_\lambda \sim \text{Gamma}(1, 20)$ .

**Cyclic power model (CPM)** Two versions of the CPM, one- and two-cycle power models (1CPM and 2CPM), were compared with the MLLM. Another version of the CPM, zero-cycle power model (0CPM), was not tested because its prediction is qualitatively similar to that of MLLM, likely making the models indistinguishable. The MLLM would not predict the power function with the value of the exponent (i.e.,  $\beta$ ) larger than 1, but such large  $\beta$  values in the 0CPM were rarely reported (Slusser et al., 2013; Spence, 1990). The 1CPM assumes that there are two reference points, at the two ends of the number line. The 2CPM assumes the midpoint as an additional reference point in addition to the two ends. We used the models provided by Hollands and Dyre (2000), which extended the model in Spence (1990). The model equations are as follows:

$$1CPM : y = U \left( \frac{x^\beta}{x^\beta + (U - x)^\beta} \right) \tag{4}$$

$$2CPM : y = \begin{cases} (U/2) \left( \frac{x^\beta}{x^\beta + (U/2 - x)^\beta} \right), & \text{if } x \leq U/2. \\ U/2 + (U/2) \left( \lambda \frac{x^\beta}{x^\beta + (U/2 - x)^\beta} \right), & \text{if } x > U/2. \end{cases} \tag{5}$$

In the equations above,  $y$  is the number estimate,  $x$  is the given number,  $U$  is the upper bound of the number line, and  $\beta$  is the exponent of the power function. CPMs predict linear functions with  $\beta = 1$ , and predict more cyclic biases as  $\beta$  deviates more from 1. For the 2CPM, we assumed the use of a reference point in the middle ( $U/2$ ). The models were fitted using the same procedure as the MLLM. The model parameters in the hierarchical Bayesian model for participant  $i$  were defined as follows:

$$\beta_i \sim \text{Gamma}(\mu_\beta / \eta_\beta, \eta_\beta), \text{ where } \mu_\beta \text{ and } \eta_\beta \text{ are hyperparameters following } \mu_\beta \sim \text{Gamma}(1, 1), \text{ and } \eta_\beta \sim \text{Gamma}(2, 1).$$

**Model comparison analysis** The MLLM and CPMs were fitted to the Gaussian process (GP) posterior mean functions estimated by the GPAL algorithm from the experimental data. Model fits were compared by the deviance information criterion (DIC; Spiegelhalter et al., 2002), which is a generalization of the Akaike information criterion (AIC; Akaike, 1998) for hierarchical Bayesian models. The DIC evaluates model fit while being penalized by model complexity (e.g., number of parameters). The model comparison based on the DIC supports the model with the smallest DIC value. The model comparison results are described and discussed in the following sections.

## Results and discussion

Studies that use number-line tasks typically use the medians of the raw data to recover an underlying function. This process was not necessary in the present study because GP automatically interpolated raw data to form an estimate of the underlying function. Therefore, instead of median points, we used the posterior mean of the estimated GP functions (see Appendix for outlier detection method). The GP posterior means averaged over participants are shown in Fig. 3.

Despite the small number of trials in each block (45 for adults; 25 for children), the functions inferred by GPAL were consistent between the first and the second blocks (i.e., test and retest). Reliability of GPAL was measured using the concordance correlation coefficient (CCC; Lawrence & Lin, 1989), a measure of agreement between the first and second

measurements. The mean CCC between the posterior mean functions from test and retest (Fig. 3) over participants was 0.89 for children (SD: 0.11) and 0.90 for adults (SD: 0.15). The strong correlation demonstrated high reliability of GP predictions.

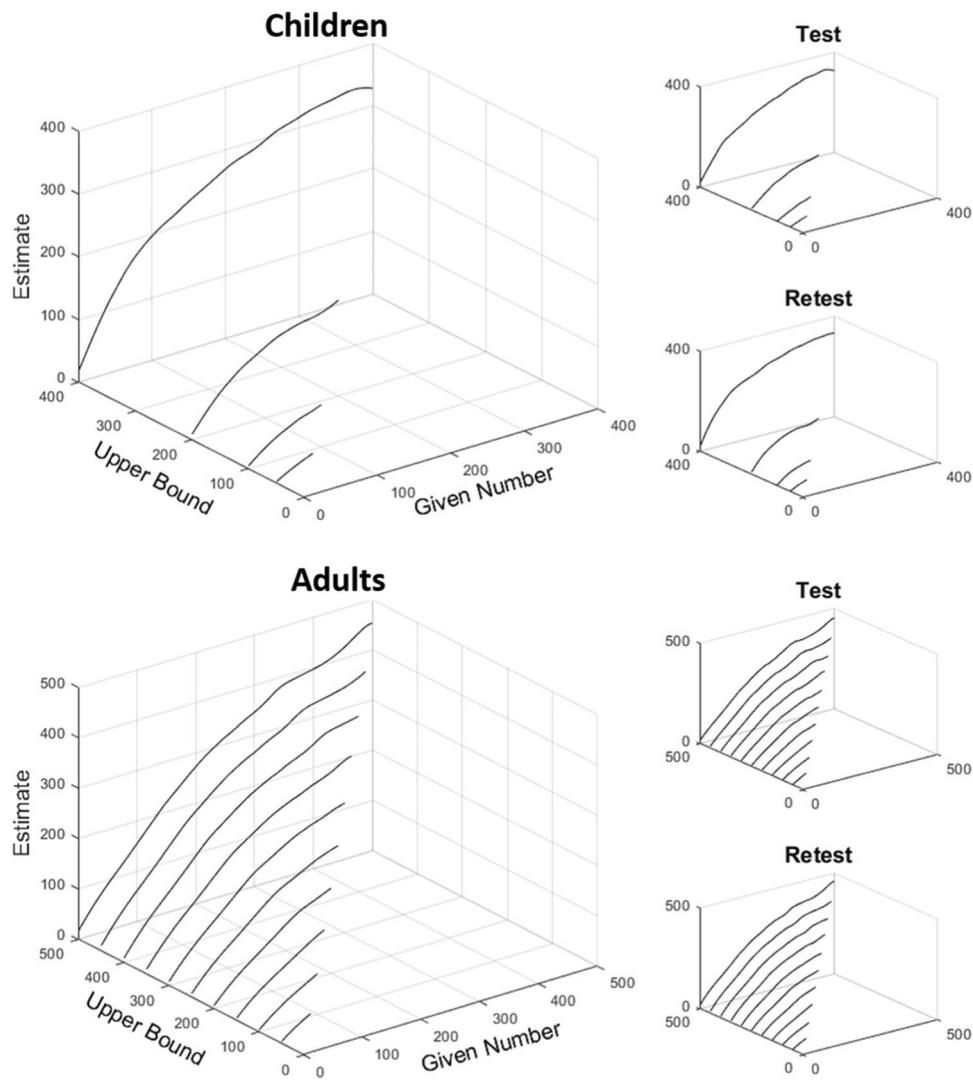
Next, we compared the MLLM and CPMs using GP-estimated posterior mean functions (see Appendix for the results of the same analysis with the raw data) to determine how we should interpret the shapes of the functions. GP functions would be better explained by the MLLM if they were from numerical representation or by the CPMs if they were from proportional reasoning skills. Table 2 shows that the DIC values were the lowest for the MLLM regardless of the upper bound, for both children and adults. A better fit of the MLLM over the CPMs suggests that non-symbolic number estimates may reflect log-linear representations for number. Therefore, we interpreted the results in a log-linear framework. Results are discussed in relation to the three questions raised in the Introduction.

### Characteristic features of number-line estimation

**Log-to-linear developmental improvement** To measure the degree of logarithmic compression in the GP functions, the posterior mean of the GP was fitted by the hierarchical Bayesian MLLM (see Appendix for the results with the raw data). The developmental changes in logarithmicity were explored by dividing participants into three age groups, younger children (kindergarteners – 2nd graders;  $n = 32$ ), older children (3rd – 7th graders;  $n = 39$ ), and adults ( $n = 159$ ). Figure 4 shows the posterior means of the hyperparameter  $\mu_\lambda$ , which represents the mean of the prior distribution for  $\lambda$ , estimated from the data combined over test and retest blocks for each age group across upper bounds. Larger values of the posterior mean indicate more logarithmic estimation.

Overall, the values of  $\mu_\lambda$  were the largest for younger children and decreased with age. Younger children’s estimation was highly logarithmic, whereas adults’ estimation was much more linear. The estimates of older children had intermediate levels of logarithmicity. The developmental decrease in  $\mu_\lambda$  from the novel multi-scale task is consistent with previous research showing that the logarithmicity components ( $\lambda$ ) decreases with age in a conventional fixed number-line task with symbolic numbers (Kim & Opfer, 2017, 2020; Opfer et al., 2016; Siegler & Booth, 2004).

Larger number ranges led to more logarithmic estimation. The posterior mean  $\mu_\lambda$  value in every age group was the smallest at the smallest upper bound of 50 (Fig. 4). In 0–50 number lines, younger children showed somewhat logarithmic estimation, with the posterior mean of  $\mu_\lambda$  being 0.12 (95% highest posterior density interval (HPDI): [0.02,0.22]; see Appendix for interpretation of HPDI intervals). Older



**Fig. 3** GP-estimated functions for children and adults. *Note.* The large plots on the left side are the mean functions collapsed across test and retest blocks. The small plots on the right side show the functions from the test and retest blocks separately

children and adults were highly linear in estimation and differed little among one another. The posterior mean of  $\mu_\lambda$  was 0.02 (95% HPDI: [0.01,0.04]) for older children, and 0.06 (95% HPDI: [0.04,0.08]) for adults. The age-related differences were much larger in 0–400 number lines. With the upper bound of 400, the posterior mean of  $\mu_\lambda$  was 0.69 (95% HPDI: [0.60,0.78]) for younger children, 0.45 (95% HPDI: [0.37,0.54]) for older children, and 0.20 (95% HPDI: [0.16,0.24]) for adults, presenting the most salient log-to-linear transition in development.

**Coexistence of log-linear representations** The changes in  $\mu_\lambda$  across the upper-bound numbers were more apparent in younger children, but estimates of older children and adults also became more logarithmic on large-scaled number lines. Even adults who were least affected by the number range

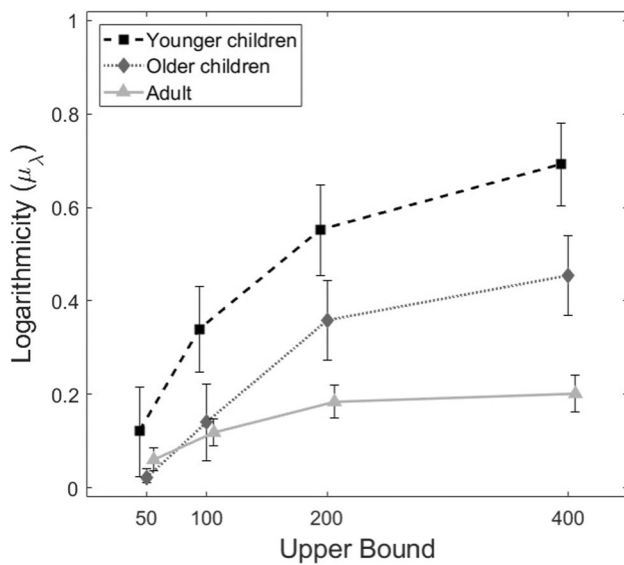
showed somewhat logarithmic estimation with a large upper-bound number. Specifically, adults’ estimation on 0–400

**Table 2** DIC values for the hierarchical Bayesian models MLLM, 1CPM, and 2CPM

Upper bound	Children			Adults		
	MLLM	1CPM	2CPM	MLLM	1CPM	2CPM
50	16518	27349	26527	29330	52750	62985
100	24659	30732	32126	45514	60995	73945
200	28014	34246	38394	57280	71550	86077
400	35494	41257	44689	75555	86677	98882
Mean	26171	33396	35434	51920	67993	80472

*MLLM* mixed log-linear model, *1CPM* one-cycle power model, *2CPM* two-cycle power model





**Fig. 4** The logarithmicity measure ( $\mu_\lambda$ ) of Gaussian process (GP) estimates for the age groups plotted against upper bounds. *Note.* Error bars indicate Bayesian 95% highest posterior density interval (HPDI)

number lines was nearly as logarithmic as younger children’s estimation on 0–50 number lines. In line with logarithmic estimates of symbolic numbers in adults (Landy et al., 2013), this outcome suggests that Western, educated adults still rely on logarithmic representations when estimating large non-symbolic numbers.

**Logarithmicity and approximate arithmetic** The degree of logarithmicity correlated with participants’ performance in the approximate addition task. To assess the correlation, we obtained individuals’ Weber fraction ( $w$ , Pica et al., 2004) in the approximate addition task. Weber fraction is an accuracy measure with numerical ratios taken into account. For example, if  $w$  is .5, participants would reliably compare 15 versus 10 dots (i.e., 50% difference) once they correctly perform arithmetic operations.

Next, we computed Bayesian partial correlations between  $\lambda$  and  $w$  values, while controlling for the effects of education levels (see Appendix for education level coding). Overall, there were positive correlations between  $\lambda$  and  $w$  across upper bounds. The posterior mean of the correlation coefficient ( $r$ ) was 0.10 for 0–50 number lines, 0.15 for 0–100 number lines, 0.18 for 0–200 number lines, and 0.19 for 0–400 number lines. The 95% HPDIs of  $r$  included zero only for 0–50 number lines (95% HPDI = [-0.04, 0.24]). These results show considerable correlation between the two non-symbolic number tasks, suggesting that the associations between number-line estimation and math skills are found with non-symbolic numbers, not only with symbolic numbers.

## Characteristics of GPAL-selected designs

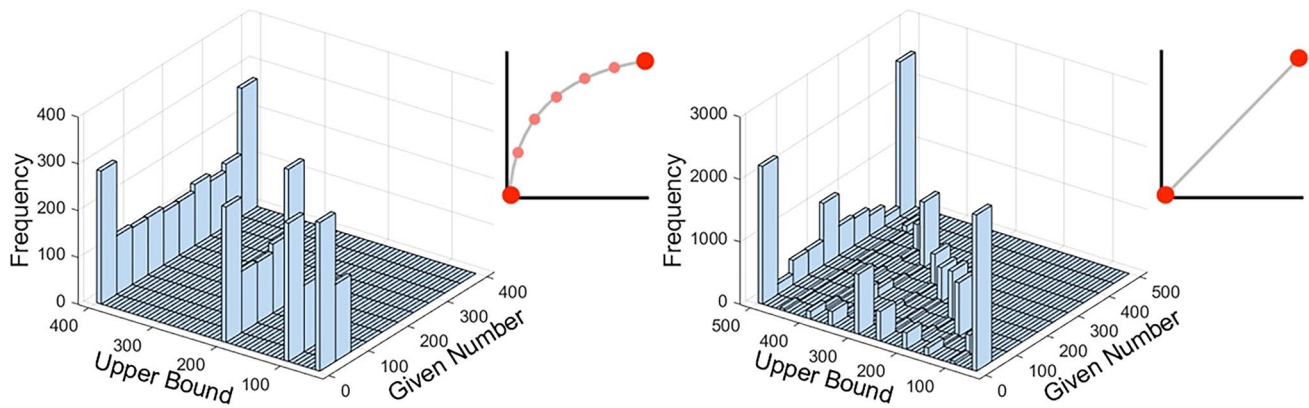
A novelty of the present experiment was that the data collection was controlled by an active learning algorithm. A combination of an upper bound and a given number was selected by GPAL trial-by-trial to optimize function estimation. The resulting frequencies of the designs selected by GPAL are shown in Fig. 5. The distribution of given numbers selected by GPAL was markedly different from typical fixed designs used in number-line tasks, where given numbers are evenly distributed over the design space (Slusser et al., 2013) or concentrated in the first half of the range (Siegler & Opfer, 2003). Rather, GPAL most frequently chose the designs at the edges of the triangular design space. This indicates that estimates of the smallest number (i.e., 5) and of the upper-bound number of each number line had the highest uncertainty (i.e., posterior variance) most of the time.

A comparison of the frequencies with which designs were selected for children and adults in Fig. 5 suggests that design selection in GPAL is sensitive to the shape of estimated function. The designs for adults were more extreme than those for children, presumably because the function was more linear for adults. Samples at the two ends of the lines are highly informative for estimating nearly straight lines in linear functions. In contrast, if the function is logarithmic, a broader sample of given numbers (i.e., designs) is required because the endpoints alone are insufficient to estimate the shape of a non-linear curve. This idea is shown in the two side plots with red circles in Fig. 5.

## General discussion

The present study examined children’s and adults’ estimates of non-symbolic numbers across multiple number ranges by adaptively optimizing the choice of design variables in a number-line task. This Bayesian active learning algorithm provided several insights into the development of non-symbolic number estimation.

The primary insight concerned the similarity of non-symbolic and symbolic number-line estimation. Our results showed all three features of symbolic number-line tasks. Specifically, we found that: (1) non-symbolic number estimation also shows a “log-to-linear shift”; (2) logarithmic and linear patterns of non-symbolic number estimates co-exist in the same individuals, with logarithmicity increasing with numeric range; and (3) logarithmicity of non-symbolic number estimates is positively correlated with the quality of estimates for arithmetic sums. The model comparison using model-neutral designs provided by GPAL supported the interpretation of these findings in the log-linear framework. All findings about non-symbolic number estimation (but the log-to-linear shift) were novel and predicted by the



**Fig. 5** Design selection frequencies in the number-line task with multiple scales *Note.* On every trial, Gaussian process active learning (GPAL) selected a given number (design variable 1) and an upper-bound number (design variable 2)

idea that symbolic and non-symbolic numbers are estimated through shared numerical representations with a common developmental pathway. Together, the results suggest that estimation of non-symbolic and symbolic numbers relies on common representations (Dehaene, 2011; Dehaene et al., 2008; Siegler & Opfer, 2003), rather than qualitatively distinct representations (Carey, 2004; Carey & Barner, 2019; Rips et al., 2008).

Another insight, which manifested in several unexpected ways, was that *size matters*. One way that it mattered was in differentiating age groups: Sensitivity to the upper bound decreased with age. Specifically, with age, the upper bound had a smaller effect on the logarithmicity of estimates, and so a large number scale (e.g., 0–400) best distinguished among age groups. This finding highlights the importance of using large numbers for studying numeric development, but it also shows the utility of GPAL in finding these numbers, which were not known a priori.

Size also mattered when addressing the association between number representations and arithmetic ability, where conclusions in the literature often conflict (Fazio et al., 2014; Halberda et al., 2008). In the case of symbolic numbers and symbolic arithmetic, there is a robust correlation (Booth & Siegler, 2008; Fazio et al., 2014; Kim & Opfer, 2017), and between non-symbolic number estimation and symbolic arithmetic, there is a weak correlation (Fazio et al., 2014). Here we looked at non-symbolic estimation and non-symbolic arithmetic, and found that logarithmicity of estimates correlated with the Weber fraction in the approximate addition task, *with the strength of the correlation tending to increase with upper bound*. These results support the idea that the approximate number system supports arithmetic intuitions (Halberda et al., 2008), but the influence of the approximate number system is reduced by crossing format (from non-symbolic to symbolic) and limiting the range of numbers tested.

Still another way that size matters is how it reveals the process of developmental change. Past studies that explored the range-dependency of the numerical estimation typically used number-line tasks with a few different scales. For example, Berteletti et al. (2010) presented 0–10 and 0–100 number lines to the same participants. Thompson and Opfer (2010) used 0–1,000, 0–10,000, and 0–100,000 number lines in a between-subject design. These studies showed that log compression increased with increasing number range in children, but the changes in logarithmicity were abrupt rather than gradual, possibly because there were only a few selected upper-bound numbers that differed by orders of magnitude. In contrast, when the upper-bound was included as a design variable that could be comprehensively controlled, we found a gradual change from more logarithmic to more linear estimation in all age groups. Thus, GPAL revealed an unexpected feature of developmental change that will be important to follow in future studies.

GPAL was particularly useful in the current study because we explored participant behaviors in a large design space in an efficient and model-free manner. Alternative methods such as using a fixed design or model-based experimental design optimization could be employed, but they would reduce the efficiency in data collection and also the informativeness of the data. That is, number-line tasks with fixed designs would require many more trials or multiple sessions to compare computational models and estimate a logarithmicity measure across upper bounds as in the current study. A model-based design optimization method (e.g., ADO; Cavagnaro et al., 2010) could facilitate parameter estimation (e.g., estimation of the logarithmicity parameter of the MLLM) across upper bounds, but the model would predefine the relation between the upper bound and logarithmicity, which may mischaracterize representation of number. When this happens (i.e., model misspecification), a design optimization algorithm might bias the design selection so

the conclusion of the study. GPAL overcomes both of these limitations, namely, inefficient data collection and model misspecification.

In conclusion, the present study explored non-symbolic number estimation in children and adults using a novel active learning algorithm (GPAL). The three characteristic features of symbolic number-line estimation were observed in the study. The results, taken together, suggest that non-symbolic and symbolic numbers may be represented through common processes that share a developmental trajectory.

## Appendix

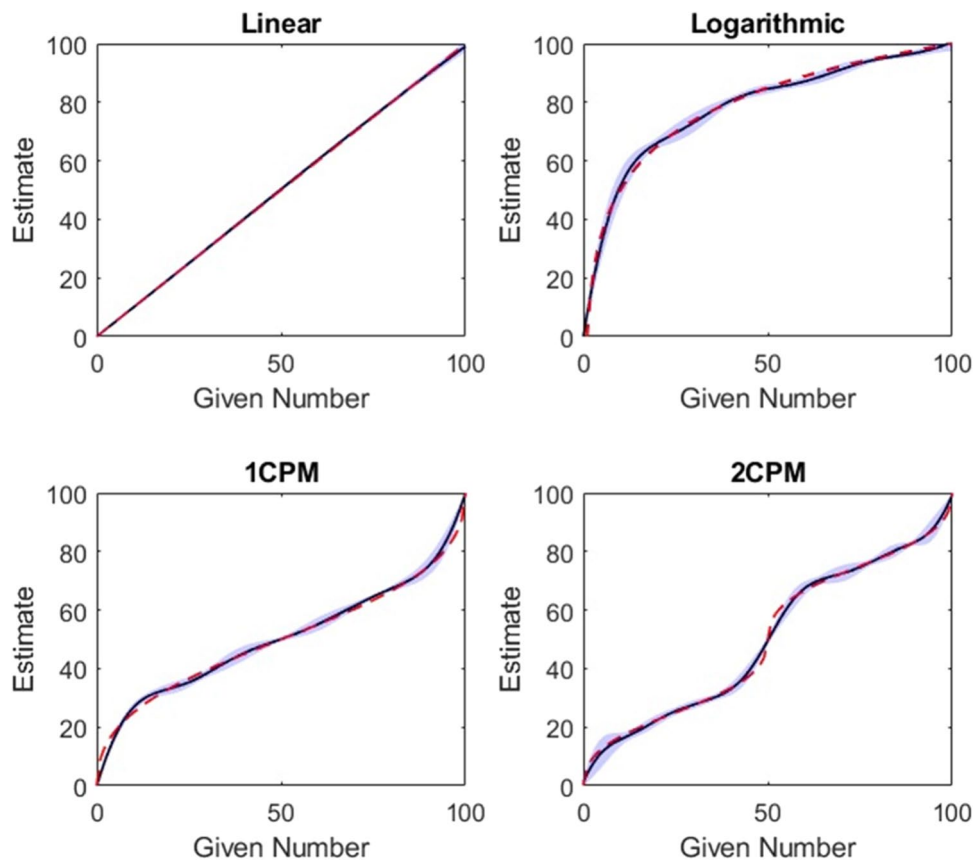
### 1. GPAL simulations

In the main body of the paper, we introduced several functions that are likely to be observed in the number-line task. Here, in simulations with artificial data, we assessed the technical soundness and ability of GPAL to identify and recover these functions as data-generating models. That is,

if GPAL works as claimed, the method should be able to recover the functional form of any model, including MLLM and CPMs.

Specifically, in the simulations, we used four different functions, which are linear (MLLM with  $\lambda = 0$ ), logarithmic (MLLM with  $\lambda = 1$ ), 1CPM ( $\beta = 0.5$ ), and 2CPM ( $\beta = 0.5$ ). The four models were then used to generate estimates for given numbers which are selected by GPAL. Normal random errors with the variance of 25 were added to the estimates. For simplicity, the number range was fixed to 0–100. This way the simulation of GPAL with ten trials was repeated 100 times for each of the four functions. Appendix Figure 6 shows results obtained at the end of ten simulated experimental trials, averaged over 100 independent simulation runs. The solid black curves in the graphs are the GPAL-inferred functions. They were practically indistinguishable from the data-generating functions (dotted red curves), thereby demonstrating that GPAL can successfully recover all the functions of interest.

### 2. Outlier detection



**Fig. 6** Results of function recovery simulations using GPAL. *Note.* In each graph, the solid black curve is the Gaussian process (GP) mean function obtained as an average over 100 independent simulation

runs. The dotted red curve is the data generating function (ground truth). The blue area depicts the 95% confidence region

Participants with extremely high posterior variance of GP (> 3SD from the mean) were excluded from the analysis, because the predictions of GP with particularly high variance were not considered reliable. Two children and six adults met this criterion. These outliers were detected separately for children and adults.

### 3. Model fitting with the raw data

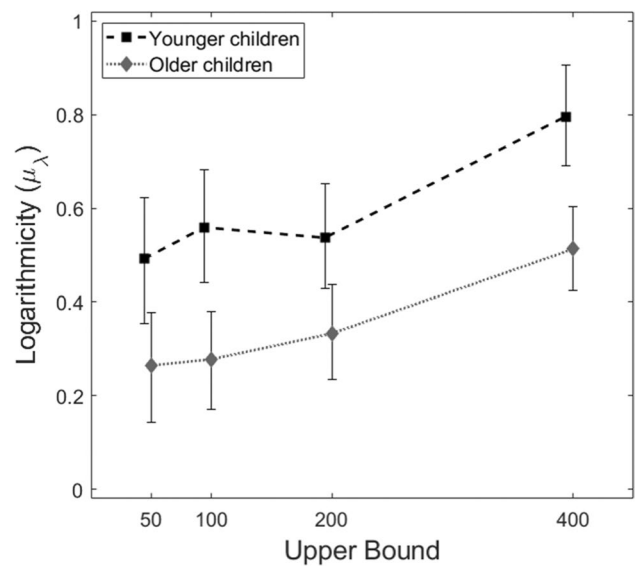
The model evaluation and parameter estimation in the current study relied on the GP-estimated posterior mean functions instead of the raw data. We also fitted the same three models of interest to the raw data to explore whether the sparse data obtained by GPAL would lead to the same outcome and conclusion. The raw data from children generally supported the results in the current study. Appendix Table 3 shows the DIC values of the MLLM and the CPMs fitted to the raw data. The MLLM generally showed smaller DIC values than the CPMs, as in the model comparison using the GP-estimated posterior mean functions (Table 2). The estimates of the logarithmicity measure ( $\mu_\lambda$ ) in the MLLM obtained from the raw data (Appendix Fig. 7) also showed the patterns consistent with those in Fig. 4. Younger children showed larger posterior mean of  $\mu_\lambda$  than older children, with increasing trends of  $\mu_\lambda$  across upper bounds in both age groups. However, large HPDIs, especially with 50 as the upper bound, suggested that the sparse raw data in each number range were not as reliable as the functions inferred by GP using the full data across number ranges. The raw data analysis was not feasible with adults because their data were extremely sparse, as shown in Fig. 5. Some adults had only one or two data points for some number ranges, making the by-range model fitting unviable.

### 4. Interpretation of the 95% highest posterior density intervals (HPDIs) in Fig. 4

**Table 3** DIC values for the hierarchical Bayesian models MLLM, 1CPM, and 2CPM. Measured with the raw data from children

Upper bound	Children		
	MLLM	1CPM	2CPM
50	5791	9797	8430
100	9623	12058	11581
200	13296	14859	15746
400	17370	18526	19977
Mean	11520	13810	13934

MLLM mixed log-linear model, 1CPM one-cycle power model, 2CPM two-cycle power model



**Fig. 7** The logarithmicity measure ( $\mu_\lambda$ ) from the raw data of children plotted against upper bounds. Note. Error bars indicate Bayesian 95% highest posterior density interval (HPDI)

In Bayesian inference, statistical evidence for the difference between the posterior distributions of  $\mu_\lambda$  from different age groups were considered strong when their 95% HPDIs did not overlap with each other. One concern about the analysis with the three age groups was that dividing children into two groups with reduced sample sizes might substantially reduce statistical power. However, HPDIs across upper bounds and age groups shown in Fig. 4 (and also Appendix Fig. 7) suggest that the divided groups have sufficiently strong statistical power to conclude that the values of  $\mu_\lambda$  are meaningfully different across upper bounds and age groups.

### 5. Education level coding in the partial correlation analysis

To control for the effects of education levels in the Bayesian partial correlations between  $\lambda$  and  $w$  values, the education level was coded as 0 for kindergarteners, as the grade for 1st–7th graders, and as 13 for adults. Thirteen corresponds to a high school graduate, which was the minimum education level of the adult participants. We did not differentiate adults by education level because their number-line estimation varied little compared to children. For the correlation analysis, the data were collapsed over children and adults, for 50, 100, 200, and 400 upper bounds.

**Funding** The present work was supported by grant FA9550-16-1-0053 to MAP and JIM from the Air Force Office of Scientific Research

(AFOSR) and R305A160295 to JEO from the Institute of Education Sciences (IES).

**Data and code availability** The data and code for all experiments and models are available for download at <https://osf.io/fvxyb/>. None of the experiments described here were preregistered.

## Declarations

**Ethics approval** Experiments in the current study were approved by Institutional Review Board (IRB) in the Ohio State University.

**Consent to participate** Informed consent was obtained from all adult participants and legal guardians of children included in the study.

**Consent for publication** Participants were informed that no identifying information about participants will be available in the article.

**Conflict of interest** There are no known conflicts of interest regarding this article.

## References

- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In: *Selected papers of hirotugu akaike* (pp. 199–213). Springer. [https://doi.org/10.1007/978-1-4612-1694-0\\_15](https://doi.org/10.1007/978-1-4612-1694-0_15)
- Anobile, G., Cicchini, G. M., & Burr, D. C. (2012). Linear mapping of numbers onto space requires attention. *Cognition*, *122*(3), 454–459. <https://doi.org/10.1016/j.cognition.2011.11.006>
- Atkinson, A., & Donev, A. (1992). *Optimum Experimental Designs*. Oxford University Press.
- Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science*, *14*(1), 125–135. <https://doi.org/10.1111/j.1467-7687.2010.00962.x>
- Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology*, *46*(2), 545–551. <https://doi.org/10.1037/a0017887>
- Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development*, *79*(4), 1016–1031. <https://doi.org/10.1111/j.1467-8624.2008.01173.x>
- Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59–68.
- Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in Cognitive Sciences*, *23*(10), 823–835. <https://doi.org/10.1016/j.tics.2019.07.004>
- Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information based approach to model discrimination in cognitive science. *Neural Computation*, *22*(4), 887–905. <https://doi.org/10.1162/neco.2009.02-09-959>
- Chang, J., Kim, J., Zhang, B.-T., Pitt, M. A., & Myung, J. I. (2021). Data-driven experimental design and model development using Gaussian Process with active learning. *Cognitive Psychology*, *125*, 000–000. <https://doi.org/10.1016/j.cogpsych.2020.101360>
- Clarke, B., Strand Cary, M. G., Shanley, L., & Sutherland, M. (2018). Exploring the promise of a number line assessment to help identify students at-risk in mathematics. *Assessment for Effective Intervention*, 151–160. <https://doi.org/10.1177/1534508418791738>
- Cohen, D. J., & Sarnecka, B. W. (2014). Children’s number-line estimation shows development of measurement skills (not number representations). *Developmental Psychology*, *50*(6), 1640–1652. <https://doi.org/10.1037/a0035901>
- Cox, G. E., Kachergis, G., & Shiffrin, R. M. (2012). Gaussian process regression for trajectory analysis. In: *Proceedings of the 34th annual conference of the cognitive science society* (pp. 1440–1445).
- Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. OUP USA.
- Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? distinct intuitions of the number scale in western and amazonian indigene cultures. *Science*, *320*(5880), 1217–1220. <https://doi.org/10.1126/science.1156540>
- Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology*, *123*, 53–72. <https://doi.org/10.1016/j.jecp.2014.01.013>
- 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>
- Geary, D. C., Hoard, M. K., Nugent, L., & Byrd-Craven, J. (2008). Development of number line representations in children with mathematical learning disability. *Developmental Neuropsychology*, *33*(3), 277–299. <https://doi.org/10.1080/87565640801982361>
- Griffiths, T. L., Lucas, C., Williams, J. J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. In *Advances in Neural Information Processing Systems*, *21*, 553–560.
- Halberda, J., Mazocco, M. M. M. & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*, 665 – 668.
- Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: the cyclical power model. *Psychological Review*, *107*(3), 500–524. <https://doi.org/10.1037/0033-295X.107.3.500>
- Honoré, N., & Noël, M. P. (2016). Improving preschoolers’ arithmetic through number magnitude training: The impact of non-symbolic and symbolic training. *PLoS ONE*, *11*(11), e0166685.
- Kim, D., & Opfer, J. E. (2017). A unified framework for bounded and unbounded numerical estimation. *Developmental Psychology*, *53*(6), 1088–1097. <https://doi.org/10.1037/dev0000305>
- Kim, D., & Opfer, J. E. (2018). Dynamics and development in number-to-space mapping. *Cognitive Psychology*, *107*, 44–66. <https://doi.org/10.1016/j.cogpsych.2018.10.001>
- Kim, D., & Opfer, J. E. (2020). Compression is evident in children’s unbounded and bounded numerical estimation: Reply to Cohen and Ray. *Developmental Psychology*, *56*(4), 853–860. <https://doi.org/10.1037/dev0000886>
- Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction*, *25*, 95–103.
- Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300. <https://doi.org/10.1016/j.tics.2010.05.001>
- Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science*, *37*(5), 775–799. <https://doi.org/10.1111/cogs.12028>
- Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*(1), 255–268. <https://doi.org/10.2307/2532051>
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lewis, D. D., & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In: *Machine learning proceedings*

- 1994 (pp. 148–156). <https://doi.org/10.1016/b978-1-55860-335-6.50026-x>
- Lyons, I. M., Ansari, D., & Beilock, S. L. (2012). Symbolic Estrangement: Evidence Against a Strong Association Between Numerical Symbols and the Quantities They Represent. *Journal of experimental psychology: General*, *141*(4), 635–641. <https://doi.org/10.1037/a0027248>
- Maertens, B., De Smedt, B., Sasanguie, D., Elen, J., & Reynvoet, B. (2016). Enhancing arithmetic in pre-schoolers with comparison or number line estimation training: Does it matter? *Learning and Instruction*, *46*, 1–11.
- Opfer, J. E., & Martens, M. A. (2012). Learning without representational change: Development of numerical estimation in individuals with williams syndrome. *Developmental Science*, *15*(6), 863–875. <https://doi.org/10.1111/j.1467-7687.2012.01187.x>
- Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*(3), 169–195. <https://doi.org/10.1016/j.cogpsych.2006.09.002>
- Opfer, J. E., Thompson, C. A., & Kim, D. (2016). Free versus anchored numerical estimation: A unified approach. *Cognition*, *149*, 11–17. <https://doi.org/10.1016/j.cognition.2015.11.015>
- Opfer, J. E., Kim, D., Young, C. J., & Marciari, F. (2019). Linear spatial-numeric associations aid memory for single numbers. *Frontiers in Psychology*, *10*, 146. <https://doi.org/10.3389/fpsyg.2019.00146>
- Opfer, J. E., Kim, D., Fazio, L. K., Zhou, X., & Siegler, R. S. (2021). Cognitive mediators of US—China differences in early symbolic arithmetic. *PLoS ONE*, *16*(8), e0255283.
- Östergren, R., & Träff, U. (2013). Early number knowledge and cognitive ability affect early arithmetic ability. *Journal of Experimental Child Psychology*, *115*(3), 405–421. <https://doi.org/10.1016/j.jecp.2013.03.007>
- Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*(5695), 499–503. <https://doi.org/10.1126/science.1102085>
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using gibbs sampling. In: *Proceedings of the 3rd international workshop on distributed statistical computing* (pp. 1–10).
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Rips, L. J., Bloomfield, A., & Asmuth, J. (2008). From numerical concepts to concepts of number. *Behavioral and Brain Sciences*, *31*(6), 623–642. <https://doi.org/10.1017/S0140525X08005566>
- Rouder, J. N., & Geary, D. C. (2014). Children's cognitive representation of the mathematical number line. *Developmental Science*, *17*(4), 525–536. <https://doi.org/10.1111/desc.12166>
- Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology*, *30*(2), 344–357. <https://doi.org/10.1111/j.2044-835X.2011.02048.x>
- Sasanguie, D., Verschaffel, L., Reynvoet, B., & Luwel, K. (2016). The development of symbolic and non-symbolic number line estimations: three developmental accounts contrasted within cross-sectional and longitudinal data. *Psychologica Belgica*, *56*(4), 382–405.
- Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16. <https://doi.org/10.1016/j.jmp.2018.03.001>
- Sella, F., Berteletti, I., Lucangeli, D., & Zorzi, M. (2015). Varieties of quantity estimation in children. *Developmental Psychology*, *51*(6), 758–770.
- Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114. <https://doi.org/10.2200/s00429ed1v01y201207aim018>
- Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, *75*(2), 428–444. <https://doi.org/10.1111/j.1467-8624.2004.00684.x>
- Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237–250. <https://doi.org/10.1111/1467-9280.02438>
- Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science*, *11*(5), 655–661. <https://doi.org/10.1111/j.1467-7687.2008.00714.x>
- Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain, and Education*, *3*(3), 143–150. <https://doi.org/10.1111/j.1751-228x.2009.01064.x>
- Slusser, E., Santiago, R., & Barth, H. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology: General*, *142*, 193–208. <https://doi.org/10.1037/0012-1649.41.6.189>
- Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 683–692. <https://doi.org/10.1037/0096-1523.16.4.683>
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical methodology)*, *64*(4), 583–639. <https://doi.org/10.1111/1467-9868.00353>
- Steyvers, M. (2011). *MATJAGS 1.3: A matlab interface for JAGS*. <https://github.com/msteyvers/matjags>
- Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and sex differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology*, *101*(1), 20–51. <https://doi.org/10.1016/j.jecp.2008.02.003>
- Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, *81*(6), 1768–1786. <https://doi.org/10.1111/j.1467-8624.2010.01509.x>
- van 't Noordende, J. E., Kroesbergen, E. H., Leseman, P. P., & Volman, M. C. J. (2021). The role of non-symbolic and symbolic skills in the development of early numerical cognition from preschool to kindergarten age. *Journal of Cognition and Development*, *22*(1), 68–83.
- Yuan, L., Prather, R., Mix, K. S., & Smith, L. B. (2020). Number representations drive number-line estimates. *Child Development*, *91*(4), e952–e967. <https://doi.org/10.1111/cdev.13333>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.