Developmental Science

Developmental Science 14:5 (2011), pp 1194–1204 DOI: 10.1111/j.1467-7687.2011.01070.x

COMMENTARY

The powers of noise-fitting: reply to Barth and Paladino

John E. Opfer,[1] Robert S. Siegler[2] and Christopher J. Young[1]

1. Department of Psychology, The Ohio State University, USA
2. Department of Psychology, Carnegie Mellon University, USA

This is a commentary on Barth and Paladino (2011).

Abstract

Barth and Paladino (2011) argue that changes in numerical representations are better modeled by a power function whose exponent gradually rises to 1 than as a shift from a logarithmic to a linear representation of numerical magnitude. However, the fit of the power function to number line estimation data may simply stem from fitting noise generated by averaging over changing proportions of logarithmic and linear estimation patterns. To evaluate this possibility, we used conventional model fitting techniques with individual as well as group average data; simulations that varied the proportion of data generated by different functions; comparisons of alternative models' prediction of new data; and microgenetic analyses of rates of change in experiments on children's learning. Both new data and individual participants' data were predicted less accurately by power functions than by logarithmic and linear functions. In microgenetic studies, changes in the best fitting power function's exponent occurred abruptly, a finding inconsistent with Barth and Paladino's interpretation that development of numerical representations reflects a gradual shift in the shape of the power function. Overall, the data support the view that change in this area entails transitions from logarithmic to linear representations of numerical magnitude.

Introduction

Studies of numerical estimation indicate that children progress from logarithmic to linear representations of numerical magnitudes. This transition has been documented during preschool for the 0–10 range, between kindergarten and second grade for the 0–100 range, between second and fourth grade for the 0–1000 range, and between third and sixth grade for the 0–10,000 and 0–100,000 ranges (see Siegler, Thompson & Opfer, 2009, for a review). The transition from logarithmic to linear representations is evident in long-term changes seen in cross-sectional studies and in short-term changes seen in microgenetic studies (Opfer & Siegler, 2007; Opfer & Thompson, 2008; Thompson & Opfer, 2008). This transition is important educationally as well as theoretically: linearity of number-line estimates correlates strongly with ability to learn solutions to unfamiliar addition problems (Booth & Siegler, 2008), numeric recall (Thompson & Siegler, 2010), number categorization (Laski & Siegler, 2007), math grades (Schneider, Grabner & Paetsch, 2009), and mathematics achievement scores (Booth & Siegler, 2006). Evidence is causal as well as correlational; randomly chosen children who play games that improve linearity of number-line estimates also improve their arithmetic learning (Siegler & Ramani, 2009).

Against this large body of evidence for representational change, Barth and Paladino (B&P) raised three key arguments. Their first was, 'Number-line tasks can *only* be properly understood as proportion judgments' (Barth & Paladino, 2011, p. 134). On this assumption, B&P adapted two models for predicting proportionality judgments (Hollands & Dyre, 2000; Spence, 1990) and applied them to number-line estimation data. Second, B&P argued that fits of their adapted power models (and likelihood ratios) provide evidence against the logarithmic-to-linear shift: 'The proportion-judgment account, making use of a single two-parameter model, provides an equally good explanation of younger children's estimates when compared to a logarithmic model, and a better explanation of older children's estimates when compared to a linear model' (p. 131). Third, although B&P found a large change in the form of the power function (nearly identical to the logarithmic-to-linear shift illustrated in Figure 1), they argued that this change did not reflect qualitative change in representations. Instead, they argued, 'Because β values near 1 correspond to near-linear relationships, there *is* developmental change toward an increasingly linear representation of numerical magnitude, but there is no evidence of a categorical shift in the type of mental numerical representation used. Rather, in addition to the other sources of change

Address for correspondence: John E. Opfer, Department of Psychology, The Ohio State University, Psychology Building 245, Columbus, OH 43210, USA; e-mail: opfer.7@osu.edu

described here, there may be a smooth developmental change in the value of this parameter' (p. 134).

In this article, we reply to all three arguments. First, we identify the arguments that we do and do not dispute. Next, we examine whether power functions with changing coefficients more accurately capture developmental changes in number-line estimation than does the hypothesized logarithmic-to-linear transition. Three types of tests – analyses of individual participants' performance, of simulated data generated by varying the proportions of logarithmic and linear functions in the data set, and of the fit of different functions to data intentionally omitted when generating the best fitting function for the remainder of the data – converge on the conclusion that the logarithmic-to-linear transition provides a more accurate account than either of the proposed power functions.

In the third and final section, we examine whether there is 'smooth developmental change' in the value of the exponent of the power function (β). To do this, we examine trial-to-trial changes in microgenetic studies in which children received feedback intended to improve their estimates. Our finding of an abrupt change in the β parameter, often literally from one trial to the next, argues strongly for a representational shift. Overall, far from providing evidence against qualitative changes in numerical magnitude representations, B&P's evidence simply illustrates that highly flexible models – such as their adapted power functions – can fit noise as well as signal.

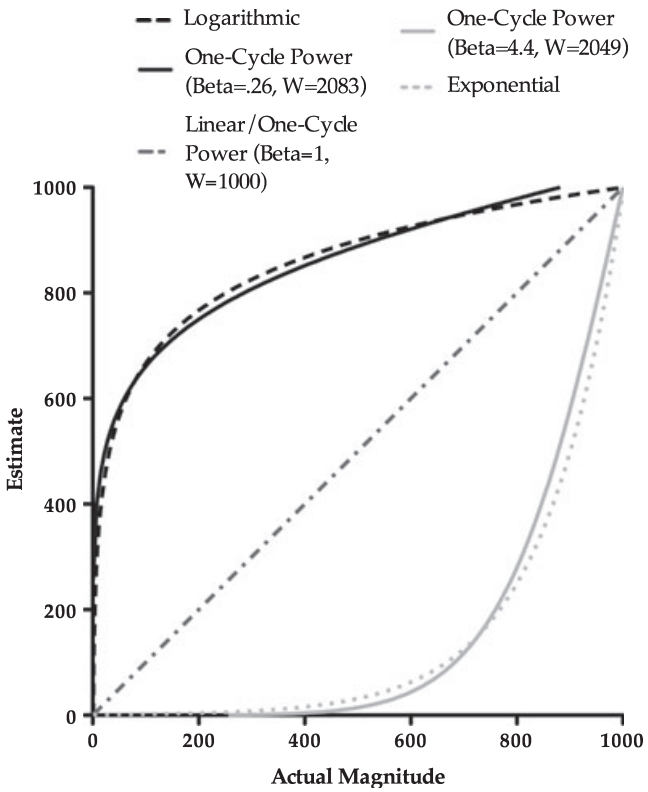

Figure 1 *Estimates predicted by ideal logarithmic, linear, and exponential functions, and three one-cycle power models that most closely fit them.*

Development of numerical magnitude representations: points of agreement and disagreement

Before replying to B&P's argument against a shift in numerical magnitude representations, we would like to highlight several points of agreement. The first is a shared conviction that findings from this area can shed light on general theoretical issues in cognitive development. These issues include the potential existence of innate representational abilities (e.g. a logarithmically compressed 'mental number line'), the extent to which early-developed capacities (e.g. non-symbolic ones) are sufficient to support development of later abilities (e.g. symbolic ones), and whether experience creates new representational resources or selects among pre-existing ones for use in novel contexts.

Another area of agreement is commitment to using mathematically explicit theories for making progress on contentious theoretical issues. Studies of numerical concepts permit a degree of mathematical precision that is much more typical of psychophysics than of studies of children's concepts in other areas (e.g. theory of mind, biology, and moral development), thereby allowing researchers to test models that generate competing quantitative predictions. From this perspective, our hypotheses have a great deal in common: both are mathematically explicit models, derived from basic research on perceptual judgments, and both use accuracy of *quantitative* predictions to address issues of qualitative importance.

A third area of agreement is the utility of the number-line task itself. Given widespread use of number-line estimation as a measure of the quality of numerical representations (e.g. Geary, Hoard, Byrd-Craven, Nugent & Numtee, 2007, 2008; Dehaene, Izard, Spelke & Pica, 2008; Muldoon, Simms, Towse, Burns & Yue, 2011; Ebersbach, Luwel, Frick, Onghena & Verschaffel, 2008; Moeller, Pixner, Kaufmann & Nuerk, 2009), agreement on this issue might seem unremarkable. However, when B&P argued that, 'Number-line tasks can *only* be properly understood as proportion judgments' (p. 134, emphasis theirs), the intended implication was not that the number line task fails to assess quality of numeric representations (Barth, personal communication). From both B&P's and our perspective, marking the location of a number on a number line requires access to some mental scaling of numerical magnitude. In this respect, the task is similar to classic psychophysical mapping tasks from Stevens (1957), who assessed perceptual experiences of brightness, loudness, line length, and other dimensions by asking participants to match the intensity of one stimulus to the intensity of another. Thus, we agree that changes in number-line estimation provide critical information for examining whether there is a shift in numerical magnitude representations.

Beyond number-line estimation, results from a wide range of numerical tasks provide a coherent picture of what early numerical magnitude representations look like: a logarithmically compressed mental number line. The hypothesis that subjective mental magnitudes increase logarithmically with actual numerical value has led to a number of accurate predictions regarding behaviors of children who are in the process of learning about these numbers. These predictions include (1) children's categorizations of numbers (e.g. as very small, small, medium, big, or very big) divide numbers such that small objective differences at the low end of the range are scattered over more categories than equal objective differences at the high end of the range (Laski & Siegler, 2007); (2) children's number bisections tend to locate the midpoint of two numbers closer to the geometric (logarithmic) mean than to the arithmetic (linear) mean (Beran, Johnson-Pynn & Ready, 2008); (3) when young children are asked to give N objects to an experimenter, the number they provide increases logarithmically with the number requested (Opfer, Thompson & Furlong, 2010); (4) estimates of the position of numbers on a line initially increase logarithmically with actual value (Siegler & Opfer, 2003); (5) estimates of the number of dots in a set where 1 and 1000 dots are provided as anchors increase logarithmically with actual value (Booth & Siegler, 2006); and (6) line lengths drawn to correspond to 1 to 1000 'zips' increase logarithmically with numerical value (Booth & Siegler, 2006). These findings follow from the Weber-Fechner Law, in which subjective differences between two quantities depend on their *proportions* (or the difference of the logarithms, which is mathematically identical).

But is the function relating objective to subjective number really logarithmic, as implied by the Weber-Fechner Law? And does the function relating objective to subjective number shift with age and experience? These are the two main issues on which we disagree.

The power function, proposed by Stevens (e.g. 1957) as a model of how people made a variety of psychophysical judgments, was the basis of the proportionality models that B&P adapted for number-line estimation. The issue of log function or power function was contentious for many years, and for good reason – depending on the exponent of the power function, power functions can be mathematically nearly identical to a logarithmic function, to a linear function, or even to an exponential function (see Figure 1).

This ability of power functions to mimic logarithmic and linear functions leads to our second disagreement. As Pitt, Myung and Zhang (2002) demonstrated, power functions are often so flexible that they can fit data (e.g. obtain high R^2 values) that were generated by the logarithmic function. They also demonstrated that the reverse is much less likely. Thus, we disagree with B&P's claim that higher R^2 values for power functions is 'evidence against a representational shift'.

Fortunately, advances in model comparison methods have provided tools for selecting among such similar functions. In the next section, we show how these advances can be applied to number-line estimation and discuss the conclusions to which they lead.

Logarithmic and linear models generate more accurate predictions

In choosing between models with similarly high R^2 values, scientists want to select models that fit signal, and they want to avoid models that fit noise. Generally, complex models – ones that have many free parameters, highly variable functional forms, and unbounded parameter values – are best at fitting noise. For this reason, complex models often fare well in conventional goodness-of-fit tests but fail to generalize to – that is, fail to predict – new data.

Which model – logarithmic or power function – is more complex? In many cases, this question can be answered by comparing the number of free parameters. From this perspective, Siegler and Opfer's (2003) logarithmic and linear models and B&P's adapted power function models (Equations 3 and 4) are equally complex. Further, B&P's non-adapted power function models (Equations 1 and 2) are less complex than the linear and logarithmic models.

However, when comparing non-linear models (e.g. power and logarithmic), number of free parameters is not the only source of complexity. Models differ in their functional form, and some functions are inherently more complex than others, making B&P's comparison of R^2 values inappropriate. A one-parameter 'black hole' model (Pitt, Myung & Zhang, 2002), for example, can fit almost any conceivable data set, but it would be absurd to prefer a 'mental number spiral' over a 'mental number line' merely because the spiral regression can account for more variance with no more parameters.

To compare the several models in question, B&P also used 'Likelihood ratios, which indicate the explanatory power of a particular model … this method allows us to compare models with different numbers of parameters' (B&P, 2011, p. 128). Assuming that their tests reflected a procedure given by their source (Glover & Dixon, 2004), Glover and Dixon recommend against B&P's practice. A similar warning has been made by others interested in model selection. Myung and Pitt (2004) offered an extensive comparison of model selection methods, and they concluded that a generalized likelihood ratio test 'is not an appropriate method for model comparison' (p. 362). Their point is especially applicable to the type of comparison we face in the present context (power vs. logarithmic functions). As Myung and Pitt note, use of likelihood ratios 'does not assess generalizability' (p. 362), 'requires a nested assumption' (p. 363) and is 'inappropriate for testing non-linear models' (p. 363) (Myung & Pitt, 2004).

Pitt and Myung (2002) provide a more useful definition of complexity: 'the property of a model that enables it to fit diverse patterns of data' (p. 422). By this definition, *a power function is more complex than a logarithmic one even when they have the same number of free parameters.* For example, as shown in Figure 1, when B&P's adapted one-cycle power model has an exponent of 1, it is identical to a linear function; when it has an exponent of .26, it is nearly identical to a logarithmic function. Thus, although their power model and our logarithmic model have the same degrees of freedom, the power model is more complex and thus risks achieving a high goodness-of-fit, due to its fitting noise rather than signal.

To assess whether this theoretical possibility fits the present case, and in particular to test whether the adapted power functions achieve a high goodness-of-fit due to noise-fitting, we examined three issues: (1) relative goodness-of-fit of the logarithmic, linear, and power models[1] to number-line estimation data, (2) impact of model complexity on fits to simulated data, where we can be absolutely sure of the functions that produced the data, and (3) generalization of the models to data not included in the original data set. (Parallel analyses of the non-adapted power functions, with similar results, are reported in the Supplemental Materials.)

Goodness-of-fit

To assess the relative fits of alternative models, we examined data from seven studies on 576 children's number-line estimates that used the standard procedure of not providing information about the location of any number except the anchors (0 and 100 or 0 and 1000): Booth and Siegler, 2006, Experiment 2; Laski and Siegler, 2007; Opfer and Siegler, 2007; Opfer and Thompson, 2008; Siegler and Booth, 2004, Experiment 1; Siegler and Laski, unpublished; and Thompson and Opfer, 2008. (Data from Siegler and Opfer, 2003, were excluded because different participants estimated values of different numbers, with too few numbers sampled for each subject for model selection procedures to be interpretable.) These studies provided eight partitions of grade and number-line task. On 0–100 number lines, we examined estimates of 72 kindergartners, 83 first graders, 77 second graders, and 13 fourth graders. On 0–1000 number lines, we examined estimates of 13 first graders, 195 second graders, 35 third graders, and 88 fourth graders.

For each partition, we examined four models – logarithmic, linear, two-cycle power function, and one-cycle power function. For each model, we regressed group median estimates for each number that was estimated against the number that was presented (see Table 1). The one-cycle power function best fit median estimates for five of eight partitions. The three exceptions were the estimates of kindergartners on 0–100 and fourth graders on 0–100 and 0–1000 (Table 1).

These results are consistent with at least two possibilities. The possibility suggested by B&P is that the data in the five sets where the one-cycle power function accounted for the most variance were generated by children whose estimates, when analyzed at the individual as well as the group level, followed a power function, with the exponent of the function increasing toward 1.00 with age and experience. The alternative possibility was that the data were generated by a mixture of two distinct sub-populations – children whose estimates fit a logarithmic function (a sub-population that became less numerous with age) and children whose estimates fit a linear function (a sub-population that became more numerous with age). As is well known, group averages sometimes lead to faulty generalizations about the cognitive processes of individual participants (Estes, 1956; Newell, 1973; Siegler, 1987).

To test whether model fits to the group medians arose from averaging over distinct sub-populations, we regressed each of the 576 individual children's estimates for each number against the number the children had been asked to estimate. We assigned a 1 to the model that best fit each participant's estimates and a 0 to the remaining models. Then we examined the association between grade and proportion of children for whom each model provided the best fit on each task. This analysis provided a more direct test than analyses of group averages for B&P's claim that 'a single two-parameter model provides an equally good explanation of younger children's estimates when compared to a logarithmic model, and a better explanation of older children's estimates when compared to a linear model'.

Table 1 *Goodness-of-fit statistics (R^2) for the models relating group median estimates against numbers to-be-estimated*

Grade	Range	Logarithmic model R^2	Linear model R^2	Two-cycle power model R^2	One-cycle power model R^2
Kind.	0–100	0.832	0.685	0.184	0.792
1st	0–100	0.933	0.921	0.733	0.967
2nd	0–100	0.851	0.95	0.866	0.967
4th	0–1000	0.825	0.997	0.993	0.996
1st	0–1000	0.688	0.63	0.396	0.73
2nd	0–1000	0.901	0.837	0.502	0.95
3rd	0–1000	0.82	0.95	0.867	0.977
4th	0–1000	0.722	0.983	0.976	0.979

[1] For the power models, we used B&P's adapted one-cycle power model, with constraints, and the original Holland and Dyre (2000) two-cycle power model. We used these forms of the models to avoid a mathematical error that is almost inevitably generated by B&P's adapted two-cycle model. Specifically, both of B&P's adapted models do not allow for W to be less than the observed range, and the adapted two-cycle model used by B&P also requires that W not be larger than the observed range. Without these constraints, the adapted two-cycle model generates imaginary predictions (e.g. the square root of a negative number) for the overestimation-underestimation pattern observed in all but one of B&P's subjects. Thus, for the types of estimates typically observed, B&P's adapted two-cycle model cannot provide a better fit than the original Holland and Dyre model.

We first compared the two-cycle power model to the logarithmic and linear models. On the 0–100 task, grade was associated with proportion of children whose estimates were best fit by the linear and logarithmic models, χ^2s (2, N = 245) = 36.74 and 35.97, ps < .0001, respectively. The linear model best fit estimates of 28% of kindergartners, 41% of first graders, 63% of second graders, and 84% of fourth graders (Figure 2). The logarithmic model best fit estimates of 71% of kindergartners, 53% of first graders, 29% of second graders, and 8% of fourth graders. Combined, the logarithmic and linear functions provided the best fit for 92–99% of children in the four grades. Less than 8% of children were best fit by the two-cycle model, a frequency that did not vary systematically with age.

On the 0–1000 task, grade was also associated with proportion of children whose estimates best fit linear and logarithmic models, χ^2s (3, N = 331) = 39.71 and 43.64, ps < .0001, respectively. The linear model best fit estimates of 15% of first graders, 32% of second graders, 54% of third graders, and 67% of fourth graders (Figure 2). The logarithmic model best fit estimates of 62% of first graders, 62% of second graders, 29% of third graders, and 14% of fourth graders. Thus, logarithmic and linear functions provided the best fit for 75–98% of individual children's estimates.

We next compared the one-cycle power model to the logarithmic and linear models. On the 0–100 task, grade was associated with the proportion of children whose estimates best fit the linear and logarithmic models, χ^2 (3 N = 245) = 11.28 and 31.09, p = .01 and p < .0001, respectively. The linear model best fit estimates of 14% of kindergartners, 23% of first graders, 30% of second graders, and 38% of fourth graders (Figure 2). The logarithmic model best fit estimates of 61% of kindergartners, 41% of first graders, 21% of second graders, and 8% of fourth graders . There was also an association between grade and proportion of children whose estimates best fit

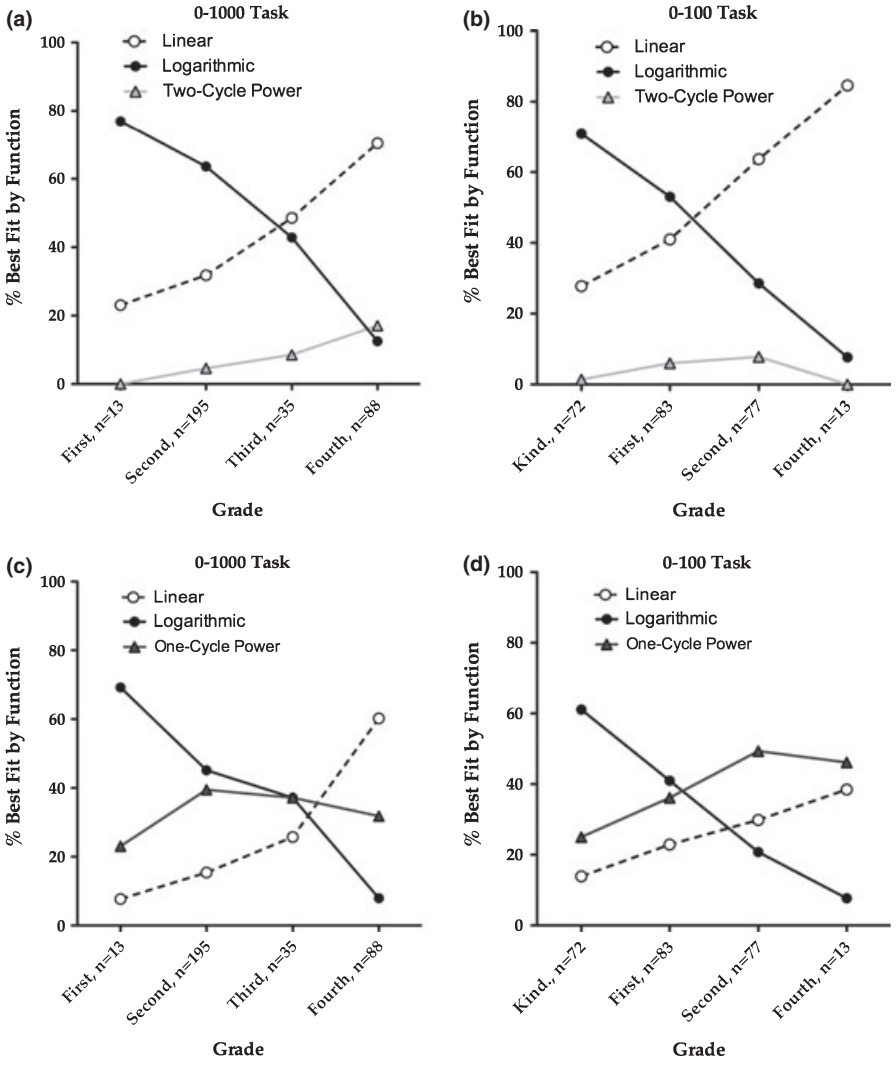

Figure 2 *Percentages of individuals in eight studies fit by each function, separated by grade, range and models compared. Upper panels compare linear, logarithmic and two-cycle power models' fits separately for (a) 0–1000 range and (b) 0–100 range. Lower panels compare linear, logarithmic and one-cycle power model fits separately for (c) 0–1000 range and (d) 0–100 range.*

the one-cycle power function, χ^2 (3 N = 245) = 10.89, $p < .05$. The power function best fit the estimates of 25% of kindergartners, 36% of first graders, 49% of second graders, and 54% of fourth graders.

On the 0–1000 task, grade was again associated with proportion of children whose estimates best fit the linear and logarithmic models, χ^2s (3, N = 331) = 20.4 and 68.1, ps < .0001, respectively. The linear model best fit the estimates of 15% of first graders, 16% of second graders, 34% of third graders, and 59% of fourth graders (Figure 2). The logarithmic model best fit the estimates of 54% of first graders, 43% of second graders, 26% of third graders, and 9% of fourth graders. Thus, the linear and logarithmic models together provided the best fit for 59–69% of individual children's estimation patterns for 0–1000 number-line estimates. The one-cycle power function provided the best fit for 23–40% of individual children's estimates and was not reliably associated with grade.

To summarize, analyses of individual children's performance indicate that the fit of the one-cycle power function to the group median estimates was largely an artifact of averaging over participants. The logarithmic and linear models usually provided the best fit to individual children's estimates.

Simulations

We hypothesized that the discrepancy between the fits of the models to the group and individual data arose from the ability of the power function to fit the kind of noisy data that arise from averaging over distinct cognitive profiles. Having seen that most analyses of individual children's data were consistent with this hypothesis, we next tested our hypothesis on simulated data in which some data were generated by logarithmic functions and the remaining data by linear functions. The advantage of this approach is that we could be 100% certain of the processes that generated these data. Thus, if power function models fit data that we knew were generated by mixes of logarithmic and linear functions, the sufficiency of our explanation of the fit of the power function model to the group medians would be demonstrated.

To be specific, we used the logarithmic function (y = 1000/ln(1000) * ln(x)) to generate 100 series of simulated estimates for all numbers 0–1000 in intervals of 25, and we used the linear function (y = x) to generate 100 series of simulated estimates for the same numbers. Then, we simulated 11 groups of 100 subjects that varied in the proportion of series that were generated by the logarithmic and the linear functions. Finally, we took the median estimates for each number from each of these 11 groups and performed conventional goodness-of-fit tests using logarithmic, linear, one-cycle power function, and two-cycle power function regression functions.

As expected, increasing the proportion of linear estimates increased the fit of the linear regression function and decreased the fit of the logarithmic regression function (Figure 3). Thus, the overall fit of the logarithmic and linear regression functions correctly mirrored the subgroups over which the data were averaged. In contrast, the fit of the one-cycle power function was uniformly high (R^2 between .90 and 1.00) *even though none of the data were generated by that or any other power function*. Thus, the one-cycle power function can generate excellent fits to group-level data generated by almost any mixture of logarithmic and linear functions. The fit of the two-cycle power function mirrored that of the linear function, again with none of the data having been generated by that power function.

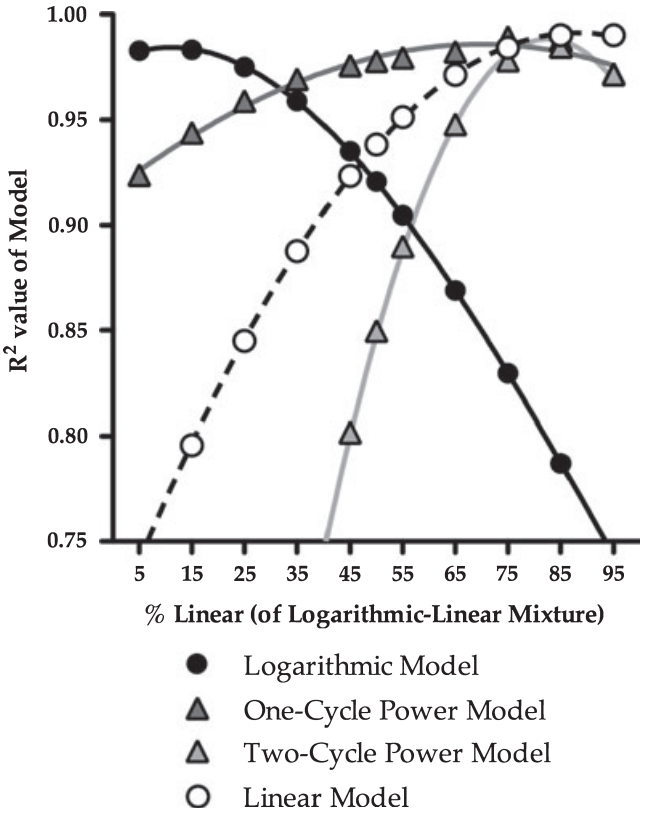

Figure 3 *Logarithmic, one-cycle power, two-cycle power, and linear models' fits to data generated by mixing 100 simulated logarithmic and linear subjects.*

We next explored whether in the actual number-line estimates of children, the power functions' fit to the averaged data arose from noise-fitting. To test this hypothesis, we aggregated data from the number-line studies we had run with 0–100 and 0–1000 tasks (see above). Participants were separated into groups according to the numerical range on which they were tested, the study from which they were drawn, and their grade. In total, these data provided us with 19 distinct groups comprising 576 children. For each group, we calculated proportion of children better fit by the linear than the logarithmic function and the R^2 value of the four regression models (logarithmic, linear, one-cycle power function, two-cycle power function) for each group.

Analyses of the data generated by the actual number-line estimates of children (Figure 4) closely mirrored analyses of the data generated by the simulation. As in

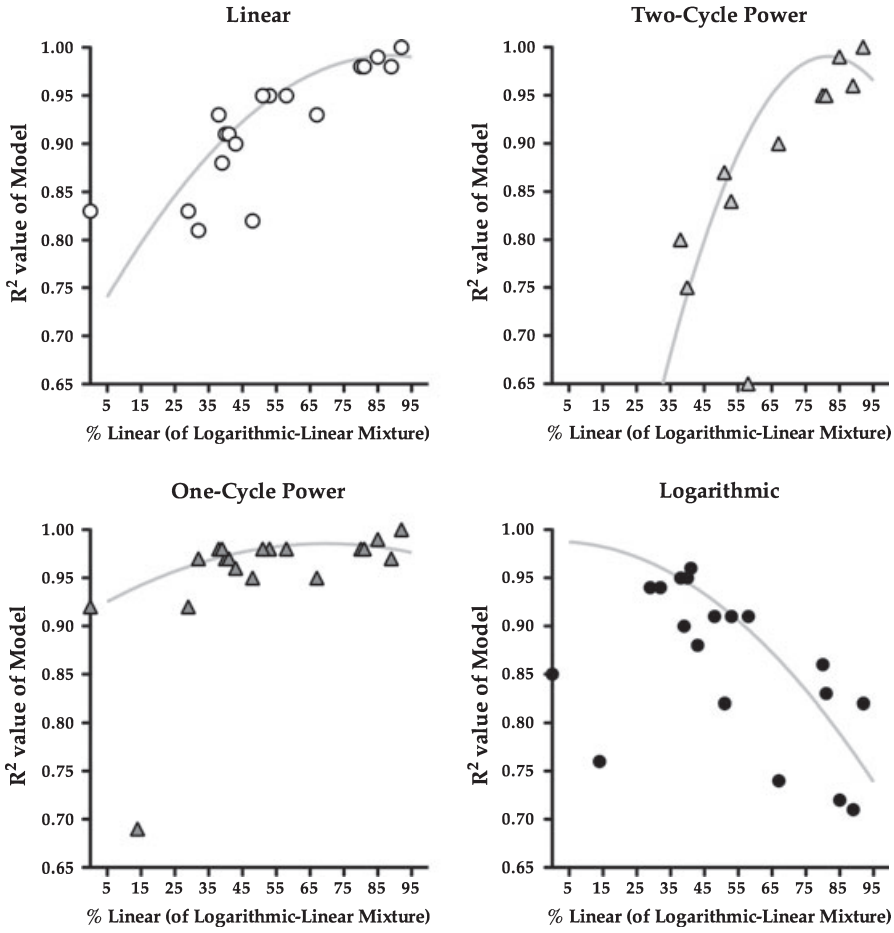

Figure 4 *Linear, two-cycle model, one-cycle power and logarithmic models' fits to median estimates of 19 grade/study/range partitions, by the percentage of subjects in each study better fit by a linear than by a logarithmic model. Each of the four models' predicted fit, based on simulated mixtures of logarithmic and linear subjects, is shown in grey curves.*

the simulation, the overall fit of the logarithmic and linear regression functions to the median estimates mirrored the relative sizes of the linear and logarithmic subgroups over which the data were averaged. In contrast, the fit of the one-cycle power function was uniformly high regardless of whether the data more closely approximated a linear or a logarithmic function (R^2 between .90 and 1), just as in the simulation where we could be sure that none of the data were generated by a power function. The fit of the two-cycle power function once again mirrored that of the linear function, just as in the simulation where none of the data were generated by a power function.

These results disconfirm B&P's contention that the fit of the power function to their data is inconsistent with a logarithmic-to-linear shift. Instead, the analyses show that the fit of the power function to group data can be explained by its excessive flexibility to fit data that were not generated by a power function.

Cross-validation of individual fits

Given the flexibility of power functions to fit noise generated by averaging over data generated by mixtures of linear and logarithmic functions, we next examined whether the four functions were equally accurate at *predicting* individual children's estimates on data that were not in the set used to generate the functions. To do this, we applied another modern model-testing technique, the leave-one-out cross-validation procedure (LOOCV; Brown, 2000). If there were no logarithmic-to-linear shift, and the fit of the power functions to the median estimates came from the functions fitting signal rather than noise, then the power functions would most accurately predict data that were not included in the original data set. In contrast, if there *is* a logarithmic-to-linear shift, and the high R^2 values of the power functions came from their ability to fit noise, then the power functions would be less accurate than the linear (or logarithmic) function for predicting omitted data.

In the LOOCV procedure, each regression model was fit to all but one data point for each participant using the ordinary-least-squares regression procedure to obtain the best-fitting parameter values. Then, the parameter values were fixed, and the model was tested on the remaining data point.

To measure accuracy of each model's predictions of the missing estimates, we first calculated the mean

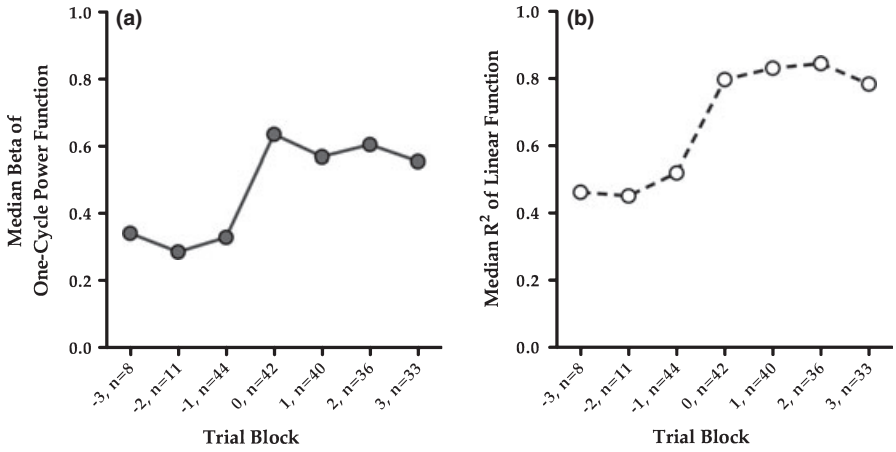

Figure 5 *Backwards trial analysis of 44 subjects in three studies assessing model fit and the β parameter of the one-cycle power model during pre-test, post-test and three feedback phases on the magnitude 150. Trial block 0 denoted the first trial block of linear estimates. Median (a) linear* R^2 *and (b) β parameter value of the one-cycle power model were collected for each trial block.*

absolute percent error (MAPE) of each model for the 576 children examined above ('goodness-of-fit'). Then, we compared the MAPEs for children who had been previously categorized as better fit by the logarithmic model or by the linear model on the basis of their estimates on all but the one problem.

For the 277 participants who were better fit by the logarithmic model than the linear one, the logarithmic model continued to yield the lowest MAPE (10.1), followed by the one-cycle power model (10.9), the linear model (13.4), and the two-cycle power model (20.7) (see Supplemental Material). For the 298 subjects whose estimates (on all but the one problem) were better fit by the linear model than the logarithmic one, the linear model yielded a lower mean MAPE (7.82) than the one-cycle power model (8.23), two-cycle power model (10.03), or logarithmic model (13.88). This patttern held true regardless of task (0–100 and 0–1000) or age group, even for combinations of task and age group where the power function provided the higher R^2 value when all data were included in the original model. Thus, results of the LOOCV converged with analyses of individual participants' estimation patterns and of the simulated data in implicating the power function models' ability to fit noise as a source of their fit to the averaged data.

Abruptness of change in estimates

B&P claimed that changes in number-line estimation are smooth and gradual (B&P, 2011, p. 134). In contrast, if children's underlying representation shifted from logarithmic to linear, the β-values in the best-fitting power functions would be expected to change abruptly.

To examine whether representations used to generate number-line estimates change gradually or abruptly, we examined trial-by-trial data from identical conditions of three different microgenetic studies of number-line estimation – Opfer and Siegler, 2007 (150-feedback condition, n = 13); Opfer and Thompson, 2008 (unpretested treatment condition, n = 15); and Thompson and Opfer, 2008 (treatment condition, n = 16). Participants in all three studies were children who generated logarithmic estimation patterns on a 0–1000 number-line pretest. After pretesting, children were presented with several blocks of number-line tests, with feedback (on the location of numbers around 150) occuring between the blocks.

To examine abruptness of changes in representations, we identified the first trial block on which the linear function provided a better fit than the logarithmic function to a given child's estimates on the no-feedback items. We labeled that set of items 'trial block 0', the immediately preceding trial block was that child's 'trial block −1', the trial block before that was the child's 'trial block −2' and so on.

These assessments made possible a backward-trials analysis of the path of change from a logarithmic to a linear representation. B&P's claim that change is gradual and continuous suggested that the fit of the linear model would gradually increase from trial block −3 to trial block +3. In this scenario, trial block 0 – the first trial block in which the linear model provided the better fit – would mark an arbitrary point along a continuum of gradual improvement, rather than the point at which children first adopted a different representation. In contrast, the hypothesis of an abrupt logarithmic-to-linear shift predicted no change in the fit of the linear model from trial block −3 to −1, a large change from trial block −1 to trial block 0, and little if any further change after trial block 0.

If children made a discontinuous shift from the logarithmic function 'y = 1000/ln[1000] * ln[x]' to the linear function 'y = x', the change in the best-fitting single-cycle power function coefficient would involve a jump from about .27 to 1 from trial block −1 to trial block 0 (see Figure 1). In contrast, if the representational change were continuous, the change in the power function coefficient should be similar for all trial blocks.

As shown in Figure 5a, the logarithmic-to-linear shift hypothesis more accurately predicted the changing values of the power function. A Kruskal-Wallis test of changes in the power function coefficients from trial block −3 to trial block −1 indicated no change in the median value of β (*p*s > .1). Similarly, no change in the coefficients was found from trial block 0 to trial block 3 (*p*s > .1). However, from trial block −1 to trial block 0, there was a large increase in the median β of the power function, from a median β = .33 to β = .64 (difference in rank sum = 72.28, $p < .001$). Thus, rather than trial block 0 reflecting an arbitrary point along 'a smooth developmental change in the value of this parameter' (B&P, 2010, p. 134), it seemed to mark the point at which children switched from a logarithmic representation to a linear one.

Changes in the linearity of estimates told a similar story (Figure 5b). From trial block −3 to −1, results of a Kruskal-Wallis test indicated no change in the median fit of the linear function (*p*s > .1). There also was no change from trial block 0 to trial block 3 (*p*s > .1). However, from trial block −1 to trial block 0, the median fit of the linear function to individual children's estimates increased from $R^2 = .52$ to $R^2 = .80$ (difference in rank sum = 48.22, $p < .01$).

Discussion

Numerous studies indicate that representations of numerical value change with age and experience: Younger children's estimates of numerical magnitude typically increase logarithmically with actual value, whereas older children's estimates increase linearly. B&P showed that this change can be modeled by a power function whose exponent gradually rises to 1, which they interpreted as evidence against a change from a logarithmic to a linear representation.

To evaluate the possibility that power models attain high goodness-of-fit values by fitting noise rather than signal, we re-examined number-line estimates from the 576 children who participated in previous studies of number-line estimation. (In our Supplemental Materials, we also re-examined data from one previous study where an anchor was provided). Like B&P, we found that one of two adapted power functions provided a good fit to median estimates of most groups of children, though the fit of the logarithmic and linear functions was better for the youngest and oldest children (respectively). We then addressed whether fits of the power model to group data provided evidence against a logarithmic-to-linear shift. Specifically, we examined whether the good fits came from the flexibility of power functions to fit extremely varied data, including logarithmic and linear patterns. The tendency of power functions to overfit data is well known among statisticians (Pitt & Myung, 2002); our concern was that this problem was present in B&P's efforts to fit the power function to number line estimation.

The powers of noise-fitting

Several different types of analyses indicated that the general problem with power functions identified by Pitt, Myung and Zhang (2002) did apply to B&P's analysis of number-line estimation. First, analyses of estimates of individual children indicated that the large majority of children were better fit by logarithmic or linear regression models than by either of the power functions used by B&P. Second, simulation models demonstrated that their adapted power models generated consistently high fits to simulated data, all of which were generated by varying proportions of logarithmic and linear functions, and none of which were generated by a power function.

We reasoned that if we were right about which children were representing numeric values logarithmically and which children were representing numeric values linearly, then we should be able to use our simulations to predict the fit of the power functions from the proportion of children who generated linear estimates. As shown in Figure 4, that test was strikingly successful: the proportion of children who were identified as using linear representations accurately predicted the fit of the one-cycle and two-cycle power functions to the group data, as well as the fit of the linear and logarithmic functions to the group data. In contrast, nothing in B&P's account indicated *when* the one-cycle or two-cycle power functions should fit the data best.

The leave-one-out cross-validation (LOOCV) analyses provided another type of evidence that the fit of the power functions was due to their extreme flexibility rather than to their accurately capturing an underlying representation. When we applied the LOOCV procedure to cases where the power function yielded their best fit, such as the second and fourth graders' estimates on the 0–100 task, the power functions were less accurate than the logarithmic and linear functions in predicting the data to which they had not been fit (Figure 5). This inability to predict new data again implies that the power of the power function came from its ability to fit noise. It also demonstrates that likelihood ratios can be inadequate for selecting among non-linear models.

Evidence for representational change

The predictive power of logarithmic and linear functions – by itself – provides limited evidence of representational change. The two functions might be good predictors of number-line estimates, but the representations generating these estimates might not change qualitatively with age and experience. If this were the case, then B&P might be right that the change in linearity of estimates was gradual and continuous.

To assess this hypothesis, we re-examined trial-to-trial data on the fit of the linear function and the β parameter. Here, again, we found strong evidence for qualitative change: Both linearity and the β parameter of the power

function increased abruptly from the last trial block on which the linear function did not provide the best fit to the first trial block where it did (Figure 5).

This finding, together with previously cited analyses of individual children's data, the simulated data, and the LOOCV tests, suggests that developmental changes in number-line estimates are not likely to stem from changing coefficients of a power function or to local repairs to children's errors following feedback. Instead, the data suggest that developmental changes in number-line estimation involve shifts in the proportions of children using logarithmic or linear representations, with the shift from logarithmic to linear functions occurring earlier for smaller numerical ranges. More generally, this example demonstrates the dangers of overfitting data with power functions, the need to test whether averaged data are consistent with individual participants' performance, the need for appropriate choices of model selection statistics, and the need for converging sources of evidence for conclusions regarding mental representations.

Acknowledgements

The research described in this paper was funded in part by grants R305A080013 and R305H050035 from the Institute of Education Sciences, in addition to support from the Teresa Heinz Chair at Carnegie Mellon University. The authors would like to thank Julie Booth, Clarissa Thompson, and Elida Laski for kindly sharing their data with us for this project. We would also like to thank Yun Tang and Jay Myung for their helpful discussions on model selection. Finally, we would like to thank Hilary Barth and Justin Halberda for their comments on an earlier draft of this manuscript.

References

Adolph, K.E., Robinson, S.R., Young, J.W., & Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychological Review*, **115**, 527–543.

Barth, H.C., & Paladino, A.M. (2011). The development of numerical estimation: evidence against a representational shift. *Developmental Science*, **14**, 125–135.

Beran, M.J., Johnson-Pynn, J.S., & Ready, C. (2008). Quantity representation in children and rhesus monkeys: linear versus logarithmic scales. *Journal of Experimental Child Psychology*, **100**, 225–233.

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology*, **4**, 545–551.

Booth, J.L., & Siegler, R.S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, **41**, 189–201.

Booth, J.L., & Siegler, R.S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development*, **79**, 1016–1031.

Brown, M.W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.

Dehaene, S. (Ed.) (1993). *Numerical cognition*. Oxford: Blackwell.

Dehaene, S., Izard, V., Spelke, E.S., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian cultures. *Science*, **320**, 1217–1220.

Ebersbach, M., Luwel, K., Frick, A., Onghena, P., & Verschaffel, L. (2008). The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9-year-old children: evidence for a segmented linear model. *Journal of Experimental Child Psychology*, **99**, 1–17.

Estes, W.K. (1956). The problem of inference from curves based on group data. *Psychological Review*, **53**, 134–140.

Franconeri, S.L., Bemis, D.K., & Alvarez, G.A. (2009). Number estimation relies on a set of segmented objects. *Cognition*, **113**, 1–13.

Geary, D.C., Hoard, M.K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development*, **78**, 1343–1359.

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**, 277–299.

Glover, S., & Dixon, P. (2004). Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, **11**, 791–806.

Hollands, J.G., & Dyre, B. (2000). Bias in proportion judgments: the cyclical power model. *Psychological Review*, **107**, 500–524.

Laski, E.V., & Siegler, R.S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development*, **76**, 1723–1743.

Moeller, K., Pixner, S., Kaufmann, L., & Nuerk, H.-C. (2009). Children's early mental number line: logarithmic or decomposed linear? *Journal of Experimental Child Psychology*, **103**, 503–515.

Muldoon, K., Simms, V., Towse, J., Burns, V., & Yue, G. (2011). Cross-cultural comparisons of 5-year-olds' estimating and mathematical ability. *Journal of Cross-Cultural Psychology*, **42**, 669–681.

Myung, I.J., & Pitt, M.A. (2004). Model comparison methods. *Methods in Enzymology*, **383**, 351–366.

Newell, A. (1973). You can't play 20 questions with nature and win. In W.G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

Opfer, J.E., & DeVries, J.M. (2008). Representational change and magnitude estimation: why young children can make more accurate salary comparisons than adults. *Cognition*, **108**, 843–849.

Opfer, J.E., & Siegler, R.S. (2004). Revisiting preschoolers' living things concept: a microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology*, **49**, 301–332.

Opfer, J.E., & Siegler, R.S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, **55**, 169–195.

Opfer, J.E., & Thompson, C.A. (2008). The trouble with transfer: insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, **79**, 790–806.

Opfer, J.E., Thompson, C.A., & Furlong, E.E. (2010). Early development of spatial-numeric associations: evidence from spatial and quantitative performance of preschoolers. *Developmental Science*, **13**, 761–771.

Pitt, M.A., & Myung, I.J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, **6**, 421–425.

Pitt, M.A., Myung, I.J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472–491.

Schneider, M., Grabner, R.H., & Paetsch, J. (2009). Mental number line, number line estimation, and mathematical school achievement: their interrelations in Grades 5 and 6. *Journal of Educational Psychology*, **101**, 359–372.

Siegler, R.S. (1987). The perils of averaging data over strategies: an example from children's addition. *Journal of Experimental Psychology: General*, **116**, 250–264.

Siegler, R.S., & Booth, J.L. (2004). Development of numerical estimation in young children. *Child Development*, **75**, 428–444.

Siegler, R.S., & Opfer, J.E. (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. *Psychological Science*, **14**, 237–243.

Siegler, R.S., & Ramani, G.B. (2009). Playing linear number board games – but not circular ones – improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology*, **101**, 545–560.

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**, 143–150.

Sophian, C., & Kailihiwa, C. (1998). Units of counting: developmental changes. *Cognitive Development*, **13**, 561–585.

Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 683–692.

Stevens, S.S. (1957). On the psychophysical law. *Psychological Review*, **64** (3), 153–181.

Thompson, C.A., & Opfer, J.E. (2008). Costs and benefits of representational change: effects of context on age and sex differences in magnitude estimation. *Journal of Experimental Child Psychology*, **101**, 20–51.

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**, 1768–1786.

Thompson, C.A., & Siegler, R.S. (2010). Linear numerical magnitude representations aid children's memory for numbers. *Psychological Science*, **21**, 1274–1281.

Wynn, K., Bloom, P., & Chiang, W.-C. (2002). Enumeration of collective entities by 5-month-old infants. *Cognition*, **83**, B55–B62.

Received: 27 September 2010
Accepted: 28 March 2011

Supporting information

Additional Supporting Information may be found in the online version of this article:

Data S1 Prediction errors of each model fit to all but one data point (using LOOCV) by grade and task.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.