Received: 24 August 2017 | Revised: 26 July 2019 | Accepted: 20 August 2019

DOI: 10.1111/desc.12897

PAPER

Developmental Science WILEY

Do infants have a sense of numerosity? A p-curve analysis of infant numerosity discrimination studies

Rachael E. Smyth[1] | Daniel Ansari[2,3]

[1]Health and Rehabilitation Sciences, Western University, London, ON, Canada

[2]Department of Psychology, Western University, London, ON, Canada

[3]Brain and Mind Institute, Western University, London, ON, Canada

Correspondence
Daniel Ansari, Department of Psychology, Brain and Mind Institute, The University of Western Ontario, Westminster Hall, Room 325, London, ON N6A 3K7, Canada.
Email: daniel.ansari@uwo.ca

Funding information
This research was funded by Natural Sciences and Engineering Research Council of Canada Discovery Grant #342192.

Abstract

Research demonstrating that infants discriminate between small (e.g., 1 vs. 3 dots) and large numerosities (e.g., 8 vs. 16 dots) is central to theories concerning the origins of human numerical abilities. To date, there has been no quantitative meta-analysis of the infant numerical competency data. Here, we quantitatively synthesize the evidential value of the available literature on infant numerosity discrimination using a meta-analytic tool called p-curve. In p-curve the distribution of available *p*-values is analyzed to determine whether the published literature examining particular hypotheses contains evidential value. p-curves demonstrated evidential value for the hypotheses that infants can discriminate between both small and large unimodal and cross-modal numerosities. However, the analyses also revealed that the published data on infants' ability to discriminate between large numerosities is less robust and statistically powered than the data on their ability to discriminate small numerosities. We argue there is a need for adequately powered replication studies to enable stronger inferences in order to use infant data to ground theories concerning the ontogenesis of numerical cognition.

KEYWORDS
approximate number system, infants, meta-analysis, numerosity discrimination, p-curve

1 | INTRODUCTION

Over the past decades, there has been significant growth in research investigating the origin of numerical cognition (for reviews see: Nieder, 2016; Nieder & Dehaene, 2009; Núñez, 2017). Within the context of this body of research, it has been posited that human and animal infants share an innate system for the representation of numerical quantity: *Human infants, as most non-human animals, appear, since the youngest age, wired up to extract the property 'number' when presented with sets of objects…* (Piazza & Eger, 2016, p. 258). This view is concisely described by Cantlon (2012) who stated that: *Primitive quantitative abilities play a role in how modern humans learn culture-specific, formal mathematical concepts (1). Preverbal children and nonhuman animals possess a primitive ability to appreciate quantities, such as the approximate number of objects in a set, without counting them verbally.* (p. 10725). Feigenson, Libertus, and Halberda (2013) reiterated this view in a review paper: *Infants, including newborns, recognize numerical changes to arrays (even controlling for non-numerical dimensions such as surface area), compare numbers of items across sensory modalities, and add and subtract approximate quantities* (e.g., Feigenson, 2011; Izard, Sann, Spelke, & Streri, 2009; McCrink & Wynn, 2004; Xu & Spelke, 2000). *Furthermore non-verbal animals including rats, fish, monkeys, and birds, also exhibit numerical representations across diverse tasks* (Brannon & Merritt, 2011; Feigenson, Dehaene, & Spelke, 2004 *for reviews*), *which suggests that representing imprecise numerical information is likely an evolved, core capacity.* (p. 74).

Critical to the theory that there exists an innate system for the representation of numerical quantity is evidence from studies testing the ability of human infants to discriminate between numerosities.

Such evidence points to the possibility that humans may be born with intuitions about numerical quantity. To specifically measure infants' numerosity discrimination abilities, researchers have predominantly utilized looking time paradigms. In these experiments, infants are typically presented with one array of dots (of a particular numerosity) repeatedly until they start to look away for a predefined period of time. At this point infants are thought to be habituated to the numerosity of the array of dots. Following habituation, infants are presented with test trials in which a novel numerosity as well as the habituated numerosity are presented. For example, infants are repeatedly presented with 16 dots (the habituation numerosity). After infants reach the criterion for habituation, they are then presented with alternating displays of 8 dots (the novel numerosity) and 16 dots (the habituated numerosity). If looking time differences exist between the novel and habituated numerosities, then infants are said to be sensitive to their numerical difference. In other words, if infants look longer at the novel compared to the habituated numerosity they are thought to have noticed a change in numerosity.

Studies utilizing this design have investigated infants' numerosity discrimination abilities with both small (e.g., 1 vs. 3) and comparatively large numerosities (e.g., 8 vs. 16) (Antell & Keating, 1983; Clearfield, 2004; Feigenson, Carey, & Spelke, 2002; Mack, 2006; Xu & Spelke, 2000). Another type of task used to measure numerosity discrimination in the visual modality is the change detection task (e.g., Libertus & Brannon, 2010). In this task, two different streams of numerosities are presented to the infants. In one stream, a non-changing numerosity is displayed to the infant (i.e., 16 dots on each trial), and in the other stream, the infant is shown changing numerosities (i.e., 8 dots on some trials and 16 dots on others). The infant's looking time at each stream is then measured.

In addition to experiments using visual displays of numerosities, numerosity discrimination has also been measured in the auditory modality by using sequences of tones. Studies investigating infants' sensitivity to numerosity in the auditory modality use a head turn preference procedure to assess discrimination. Infants are seated on a caregiver's lap or in a carrier in an enclosed space and are habituated to sequences of a certain number of sounds. During the test phase, infants hear trials of sequences that contain the number of sounds that were played during the habituation phase as well as sequences containing a novel number of sounds. The amount of time that infants spend turning their head toward the source of novel versus familiar number of sounds is compared. If infants turn their heads longer toward the novel numerosity compared to the habituation numerosity, this is thought to indicate successful discrimination between the two numerosities (Lipton & Spelke, 2003, 2004; van-Marle & Wynn, 2009).

Another way to investigate numerosity discrimination is to use cross-modal paradigms. In cross-modal paradigms, two modalities of presentation are used (e.g., auditory and visual, or visual and tactile) either concurrently or consecutively. For example, an infant might be familiarized to tones and objects concurrently and then in the test phase hear either 4 tones or 12 tones and then be shown a congruent number of objects or an incongruent number of objects. Looking time would be measured and compared for congruent versus incongruent trials (i.e., Feigenson, 2011). Published cross-modal paradigms vary significantly in the modalities used, whether presentation occurs concurrently or consecutively, and whether habituation and test phases both include both modalities or only one of the modalities being tested. Another consideration that can vary in cross-modal numerosity matching studies is that, in some cases, infants look longer at incongruent trials (Kobayashi, Hiraki, & Hasegawa, 2005; Moore, Benenson, Reznick, Peterson, & Kagan, 1987; Starkey, Spelke, & Gelman, 1983), whereas in other experiments, infants spend more time looking at trials that are congruent (Jordan & Brannon, 2006). Depending on paradigm specifications, different behavior is expected, and supports infants' ability to discriminate between numerosities. In other words, there is significant variability in the design and dependent variables of interest among the cross-modal paradigms used to assess infants' abstract representations of numerosity.

Using all of these methods, studies have shown that human infants appear to be sensitive to numerosity differences in small sets (e.g., 1 vs. 3) and large sets (e.g., 8 vs. 16). As is evident from the quotes above, the support that these studies provide for a sensitivity to numerosity in human infants represents a critical litmus test of the notion that humans are born with a sense of numerosity.

In addition, there exists evidence to suggest that different systems of representation may be underlying infants' small and large numerosity discrimination abilities. Specifically, when infants' abilities to discriminate between small and large numerosities (e.g., when one small set is compared to one large set, such as 3 vs. 6) have been examined, infants often fail to discriminate between small and large numerosities (Feigenson, Carey, & Hauser, 2002; Xu, 2003).

The finding that infants can discriminate between each of small and large numerosities, but have difficulties discriminating between small and large numerosities (Xu, 2003) has led to the suggestion that different mechanisms underlie the discrimination of large numerosities and small numerosities (Feigenson et al., 2004; Xu, 2003). It has been posited that infants have two innate systems for numerosity discrimination: (a) an approximate system for the representation of large numerosities and (b) a precise system for the representation of 1–3 objects (Feigenson et al., 2004). The two systems are present in infancy and are thought to remain functional into adulthood (Ansari,

Research Highlights

- p-curve analysis was used to assess the evidential value of published infant numerosity discrimination studies.
- Analyses demonstrated evidential value in small numerosity discrimination data and weaker evidential value for large numerosity discrimination data.
- Replication studies are recommended to provide more data to fully establish the evidential value of infant numerosity discrimination studies.

Lyons, Van Eimeren, & Xu, 2007; Hyde, 2011; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008). The evidence for small and large numerosity processing in infants and the resulting theoretical models lie at the core of current theories concerning the origins of numerical processing (Feigenson et al., 2004; Hyde, 2011).

Given the key role that evidence from studies with human infants plays in current theories of numerosity processing and the origins of our numerical abilities, it is not surprising that key studies examining numerosity discrimination in infants have been widely cited. According to Google Scholar, as of the date of submission of the present manuscript,the first paper reporting small numerosity discrimination by Starkey and Cooper (1980) has been cited 778 times, whereas the first paper reporting large numerosity discrimination by Xu and Spelke (2000) has been cited 1,213 times.

While several qualitative reviews of the infant numerical processing literature exist (i.e., Brez, 2009; Cantrell & Smith, 2013; Feigenson et al., 2004; Libertus, Starr, & Brannon, 2014), no quantitative synthesis of the infant numerosities discrimination literature has to date been performed, and as such, the evidential value of the literature examining infant numerosity discrimination has not been thoroughly evaluated. Given the importance of the evidential value of infant numerosity discrimination data for theories of numerical cognition, it is important to evaluate the strength of the evidence on which so much current theory is built. There is certainly evidence for small and large numerosity discrimination in other species and in adults, but this study focused on human infants because of the influence of this literature on numerical cognition theories (Feigenson et al., 2004; Hyde, 2011). Therefore, in this paper we used a meta-analytic tool to evaluate the evidential value of the published data on infants' small and large numerosity discrimination abilities.

One particularly pernicious difficulty with meta-analysis is that publication bias and the 'file drawer problem' (studies reporting significant results are more likely to be published than studies reporting non-significant ones), which is very prevalent, can lead to the overestimation of effect sizes in meta-analytic estimates (Ioannidis, 2005; Simonsohn, Nelson, & Simmons, 2014a). However, a recently developed meta-analytic tool, the p-curve analysis (http://www.p-curve.com) allows for the assessment of the evidential value of a particular set of data using only the published data, whereas at the same time providing a control for publication bias. Specifically, p-curve analyses overcome the problems associated with publication bias in meta-analytic summaries of existing data by investigating the distribution of significant *p*-values to determine whether the findings of a study reflect a true result (and thus have evidential value) or whether they are the result of the file drawer problem (Simonsohn et al., 2014a; Simonsohn, Nelson, & Simmons, 2014b).

Importantly, p-curve analysis plots and analyzes the distribution of reported, significant *p*-values. Non-significant *p*-values cannot be included in a p-curve analysis. It is important to note that there are differences between small numerosity discrimination and large numerosity discrimination in the number of published non-significant results. There are more studies assessing small than large numerosity discrimination that show that infants do not discriminate between presented numerosities. Although this discrepancy has been observed, due to challenges such as the file drawer problem (non-significant results are less likely to be published than significant ones), it cannot be concluded that similar results do not exist for large numerosity discrimination (we simply do not know the number of non-significant results that have not been published).

The p-curve analysis investigates the distribution of *p*-values under the null hypothesis. *p*-values are uniformly distributed under the null hypothesis. As described in Simonsohn et al. (2014a), a *p*-value < .05 suggests that 5% of *p*-values will be .05 or less and a *p*-value of <.04 suggests that 4% of *p*-values will be .04 or less. This does not mean that 5% of *p*-values will be .05, but that 5% will be .05 and lower. The difference between .05 and .04 is therefore 1%, and it follows that the difference between each of .01, .02, .03, .04, and .05 is 1%. Under the null, 5% of results will fall below .05, which then means that of the 5% of null findings with *p*-values < .05, the *p*-values < .05 will be equally distributed. If no real effect is found, the likelihood of obtaining any *p*-value < .05 is the same. We would not expect a peak at any one p-value, but an equal likelihood of obtaining any *p*-value < .05. Inferences derived from the distribution of *p*-values are based on this reasoning. In other words, if the null hypothesis is true then observing a *p*-value of .05 is as likely as observing a *p*-value of .001 (marked by the broken red line in the p-curve output figure, see Figures 1–4). If, however, there is evidence to reject the null hypothesis, the distribution of *p*-values will be right skewed. In other words, there will be more *p*-values closer to or less than, for example, $p = .01$ than there are *p*-values closer to $p = .05$. This occurs to varying degrees depending on the size of the effect and the size of the sample. In an extreme case (e.g., a reliably large effect in a large sample), the p-curve is more likely demonstrate stronger evidence ($p < .01$). In a more moderate case, the p-curve is likely to fall somewhere between the extreme case and the null hypothesis (equal distribution). Put differently, the further the weight of *p*-values is away from the statistical significant cut off of $p < .05$, the larger the evidential value to allow for the rejection of the null hypothesis.

If, however, the distribution of *p*-values in a given body of studies investigating a particular hypothesis is left skewed, this indicates publication bias, as it suggests that more studies are published that are close to the cut-off of .05. Such left skew can result from a practice known as 'p-hacking'. In p-hacking, researchers run various analyses and publish only those that are statistically significant once they pass the commonly used cut-off of $p < .05$. If such practice is systematic in a given body of literature, a left skewed p-curve will be observed. Taken together p-curves that have a right skew (i.e., more low *p*-values [i.e., $p = .01$] than high *p*-values [i.e., $p = .04$]) provide support for evidential value of the hypothesis in question. A left skewed p-curve is more likely to be the result of publication bias and research practices such as p-hacking combined with the file-drawer problem described above: that is, the selective reporting of significant results (Simonsohn et al., 2014a).

The p-curve addresses the degree of right skew in two ways: the half p-curve and the full p-curve. The half p-curve assesses the *p*-values that fall < .025, whereas the full p-curve assesses the *p*-values

that fall across the entire $p < .05$ spectrum. The half p-curve provides a more robust analysis against p-hacking (Simonsohn, Simmons, & Nelson, 2015). The half p-curve is more robust against p-hacking because it assesses the distribution of *p*-values < .025. In doing so, the half p-curve is less likely to mistake p-hacking for evidential value (Simonsohn et al., 2015). Lakens (2014) nicely summarizes the value of p-curve analysis in controlling for publication bias by saying, *traditional meta-analyses are one approach, but suffer from publication bias. p-curve analysis is a recently developed meta-analytic procedure* (Simonsohn et al., 2014a) *that is unaffected by publication bias. p-curve analysis can differentiate between published findings that should increase our prior belief that a specific hypothesis is true, and findings that should not increase our prior belief that a specific hypothesis is true.* (p. 3)

Furthermore, p-curve analyses are ideally suited to compare the evidential value of two different hypotheses because p-curves are hypothesis driven and p-curving plots the distribution of *p*-values to assess whether the available literature suggests evidential value for each hypothesis in question (Lakens, 2014). Therefore, in the case of infant numerosity discrimination data, p-curve analyses can assess the value of the two key hypotheses: (a) infants can discriminate between small numerosities, and (b) infants can discriminate between large numerosities. Furthermore, p-curve analyses provide an analysis of the statistical power of a body of studies, while simultaneously controlling for publication bias. To assess statistical power, p-curve analysis assesses the distribution of *p*-values in the available literature and compares that to the expected distribution of *p*-values at various levels of power. A p-curve that is right skewed indicates greater power of the underlying literature, whereas a less right-skewed p-curve is indicative of lower power (Simonsohn et al., 2014a). The p-curve reduces the reliance on arbitrary assumptions in calculating power (i.e., effect size assumptions, calculations affected by publication bias) because the distribution of the curve itself allows one to make inferences about the effect size and power of the data (Simonsohn et al., 2014b).

p-curve analysis allows researchers to use the significant results reported in the literature to carry out a meta-analysis without fear of falling victim to the file drawer problem. (Simonsohn et al., 2014a, 2014b). Researchers can also use p-curve analyses to inform the direction their research takes: *Researchers may use p-curve to decide which literatures to build on or which studies to attempt costly replications of* (Simonsohn et al., 2014a, p. 535). In this way, p-curve analysis provides an excellent tool to assess the strength of evidence for key hypotheses that form the basis of influential theories, such as the notion that infants are born with a sense of numerosity, which provides the scaffold for their development of numerical and mathematical abilities.

2 | METHOD

Literature searches were performed in SCOPUS, PsycINFO, Proquest Dissertations and Theses, and ERIC using combinations of the various search terms: 'infant numerosity', 'infant approximate number system', 'infan*', 'number discrimination', and 'number sense'. Additionally, for each hypothesis we examined which papers cited the original empirical investigation of that hypothesis: Starkey and Cooper (1980) for the small numerosity hypothesis and Xu and Spelke (2000) for the large numerosity hypothesis. Finally, the literature reviews from two studies, which provided comprehensive literature reviews of the hypotheses in question were reviewed and any studies that had not been located in the previous literature searches were included in the analysis (Brez, 2009; Cantrell & Smith, 2013).

In following the guidelines laid out by Simonsohn et al. (2014a), we created an a priori inclusion rule and decided which studies to include based on the inclusion rule. To be included in this analysis, studies needed to:

1. Be investigating numerical discrimination abilities in infants;
2. Use visual stimuli, auditory stimuli or cross-modal stimuli;
3. Measure numerical discrimination using looking time paradigms, not measurements of neural activity; and
4. Compare large numerosities (four and above) OR small numerosities (1–3) which has been described as within the subitizing range (Revkin et al., 2008).

Articles were excluded if they investigated hypotheses other than those investigating numerical discrimination in infants, as is protocol in p-curve analyses. Excluded articles were also those that investigated hypotheses which may have been theoretically based on the skills required to discriminate numerosities, but also encompassed more advanced skills, such as infant arithmetic tasks. While success on arithmetic tasks may indicate an infant's ability to discriminate numerosities, a lack of success does not necessarily indicate an inability to discriminate numerosities. This type of task was not included because it did not directly or clearly address the hypothesis of interest. In addition, a meta-analysis specifically examining the arithmetic abilities of infants was recently performed and addresses the hypothesis that infants' possess rudimentary arithmetic abilities (Christodoulou, Lac, & Moore, 2017).

2.1 | Main p-curves

Four main p-curve analyses were run. The small unimodal numerosity discrimination p-curve analysis included 12 *p*-values from 11 experiments from 11 published articles. The small cross-modal numerosity matching p-curve analysis included five *p*-values from five experiments from three published articles. The large unimodal numerosity discrimination p-curve analysis included 23 *p*-values from 21 experiments from 15 published articles. The large cross-modal numerosity matching p-curve analysis included six *p*-values from five experiments from four published articles[1]. The p-curve application on www.p-curve.com calculates the exact *p*-value based on the reported statistical results from each study, and as such, if a *p*-value has been rounded down to $p < .05$ in a study, it will be excluded from a p-curve when it is calculated.

In some cases, one article reported multiple *p*-values. Because dependent *p*-values cannot be included in the analysis, there were strict inclusion criteria for multiple *p*-values from one article. Multiple *p*-values could be included if: (a) article included multiple experiments which reported *p*-values for independent samples or (b) article or individual experiment reported *p*-values for independent samples or numerosity ratios. For example, Libertus and Brannon (2010) included three experiments in their article, Stable Individual Differences in Number Discrimination in Infancy. In Experiment 1, infants were randomly assigned to one of five conditions. Because the sample in each condition was independent (completely unique), the *p*-values from each significant condition were included in the analysis, in this case three separate *p*-values. It should be noted that the disclosure tables also display the calculated *p*-value for each analysis according to Simonsohn et al. (2014a), and therefore, the disclosure tables demonstrate the precise *p*-value as opposed to the *p*-value reported in each study. The p-curve disclosure table for small unimodal numerosity discrimination can be found on Open Science Framework at: https://osf.io/4q2v3/. The p-curve disclosure table for small cross-modal numerosity discrimination can be found at: https://osf.io/smf54/. The p-curve disclosure table for large unimodal numerosity discrimination can be found at: https://osf.io/ahk4e/. The p-curve disclosure table for large cross-modal numerosity discrimination can be found at: https://osf.io/w8fsv/.

2.2 | Robustness analyses

When experiments or conditions report multiple, dependent, significant *p*-values, Simonsohn et al. (2014a) recommend performing a robustness analysis, in which the *p*-values chosen for the analysis are replaced by the dependent *p*-values, when available, and rerunning the p-curve analysis with the alternate *p*-values. The purpose of robustness analyses is to investigate the consistency of evidential value in the cases where multiple, dependent *p*-values could have qualified for the analysis. To be consistent, the first *p*-value that occurred in the paper or experiment was used in the original analysis based on the example provided in Simonsohn et al. (2014a). The robustness analyses were then run by replacing the original *p*-values from any study that had multiple qualified *p*-values with the second qualified p-value. The robustness analyses account for the situations in which multiple *p*-values from a study may address a hypothesis and is designed to ensure that evidential value exists when each of those *p*-values is included in the analysis.

2.3 | Estimation of the underlying power

Estimates of the underlying power were also run in the default p-curve analysis on www.p-curve.com and results are reported below. To obtain a power analysis, the observed p-curve is compared to multiple expected p-curves at various values of power and the power value that corresponds to the most comparable expected p-curve is selected (Simonsohn et al., 2014a).

3 | RESULTS

3.1 | Main p-curves

p-curve analyses were run on four sets of *p*-values extracted from the literature: small unimodal numerosity discrimination, small cross-modal numerosity matching, large unimodal numerosity discrimination and large cross-modal numerosity matching. A list of the *p*-values included in each analysis is provided: https://osf.io/byvg8/. As detailed in Simonsohn et al. (2014a) and Simonsohn et al. (2015), in order for the analysis to indicate evidential value, one of two adequacy conditions must be met. The full p-curve test looks at all *p*-values < .05. A significant result of a full p-curve test indicates that the *p*-values < .05 are significantly right-skewed. The half p-curve test assesses all *p*-values < .025. Again, the distribution of these *p*-values is assessed, and a significant result suggests that the *p*-values < .025 are significantly right skewed. The full p-curve test is less conservative than the half p-curve test and, alone, it cannot always distinguish evidential value from ambitious p-hacking. In response to concerns about the full p-curve test's fallibility, the half p-curve test was implemented and is run concurrently as a more thorough assessment of evidential value. There are two adequacy conditions that indicate evidential value: (a) the half p-curve test must be right-skewed with $p < .05$ or (b) the half p-curve test and the full p-curve test must both be right-skewed with $p < .1$. At least one of these conditions must be met to demonstrate evidential value. The

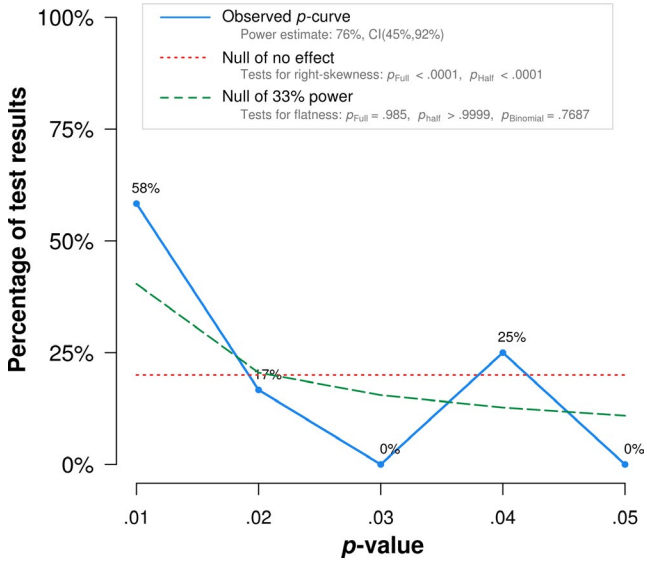

FIGURE 1 Small unimodal numerosity discrimination p-curve. The blue line shows the distribution of *p*-values from the data. The analysis is run on *p*-values in a continuous manner. If the null hypothesis of zero effect is true, the distribution of *p*-values < .05 will be equal (red line). If the null hypothesis is true and the studies lack evidential value, the p-curve will be flatter than 33% (green line). If there is evidence of p-hacking, the p-curve will be left-skewed

significance of these two tests is determined and based on the significance levels of the two distributions, the evidential value of the p-curve is assessed. In contrast, evidential value is considered inadequate or absent if one of the two inadequacy conditions are met. The two inadequacy conditions indicate an absence of evidential value: (a) if the 33% power test for the full p-curve is $p < .05$ (that is the p-curve is significantly flatter than would be the case if the average power of included studies was 33%) or (b) if both the half p-curve and binomial 33% power test are $p < .1$ (the 33% power test is marked by the dashed green line in the p-curve output figure, see Figures 1–4). A p-curve can only meet adequacy conditions or inadequacy conditions and not both, although it is possible for a p-curve to meet neither adequacy conditions nor inadequacy conditions.

3.1.1 | Small numerosity discrimination (unimodal)

The small unimodal numerosity discrimination p-curve analysis consists of 12 statistically significant ($p < .05$) results (https://osf.io/gw5jk/). Both adequacy conditions are met for the small unimodal numerosity discrimination p-curve analysis. The half p-curve ($p < .025$) is significant at a level of $p < .05$ and the full p-curve ($p < .05$) and half p-curve are both significant at a level of $p < .1$. These findings together indicate evidential value. Neither inadequacy condition is met, so the p-curve does not indicate absent or inadequate evidential value. The estimated power of the studies that comprise the p-curve is 76% (CI = 4%–92%). Figure 1 depicts the distribution of *p*-values in the small numerosity discrimination p-curve analysis.

3.1.2 | Small numerosity discrimination (cross-modal)

The small cross-modal numerosity discrimination p-curve analysis consists of five statistically significant results (https://osf.io/3wua5/). Again, both adequacy conditions are met with the half p-curve showing significance at a level of $p < .05$ and the half p-curve and the full p-curve showing significance at a level of $p < .1$, which suggests that the studies contain evidential value. Neither inadequacy condition is met, so the p-curve does not indicate absent or inadequate evidential value.

The estimated power of the small cross-modal numerosity discrimination p-curve is 96% (CI = 76%–99%). Figure 2 depicts the distribution of *p*-values in the small cross-modal numerosity discrimination p-curve analysis.

3.1.3 | Large numerosity discrimination (unimodal)

The large unimodal numerosity discrimination p-curve analysis is made up of 23 statistically significant results (https://osf.io/scvkz/). In the large unimodal numerosity discrimination p-curve, only one of the two adequacy conditions is met. The half p-curve is not significant at a level of $p < .05$, but the full p-curve and the half p-curve are both significant at the level of $p < .1$, which indicates the data contain evidential value. Importantly, neither inadequacy condition is met, so the p-curve does not indicate absent or inadequate evidential value. The estimated power of the studies that comprise the p-curve is 51% (CI = 24%–74%). Figure 3 displays the distribution of *p*-values for the large unimodal numerosity discrimination p-curve analysis.

3.1.4 | Large numerosity discrimination (cross-modal)

For the p-curve analysis of the large cross-modal numerosity discrimination findings, six statistically significant results are considered (https://osf.io/u36dp/). In the large cross-modal numerosity discrimination p-curve, both adequacy conditions are met. The full p-curve for this analysis is significantly right skewed ($p < .1$), and because the half p-curve is significantly right skewed ($p < .05$), both conditions are met. This suggests that the p-curve does indicate evidential value. Neither inadequacy condition is met, so the p-curve also does not indicate absent or inadequate evidential value. The estimated power of the studies that comprise the large cross-modal numerosity discrimination p-curve is 83% (CI = 40%–97%). Figure 4 displays the distribution of *p*-values for the large cross-modal numerosity discrimination p-curve analysis.

3.2 | Cumulative p-curve analysis

The p-curve analysis (www.p-curve.com) also reports the results of a cumulative meta-analysis, which involves progressively dropping the

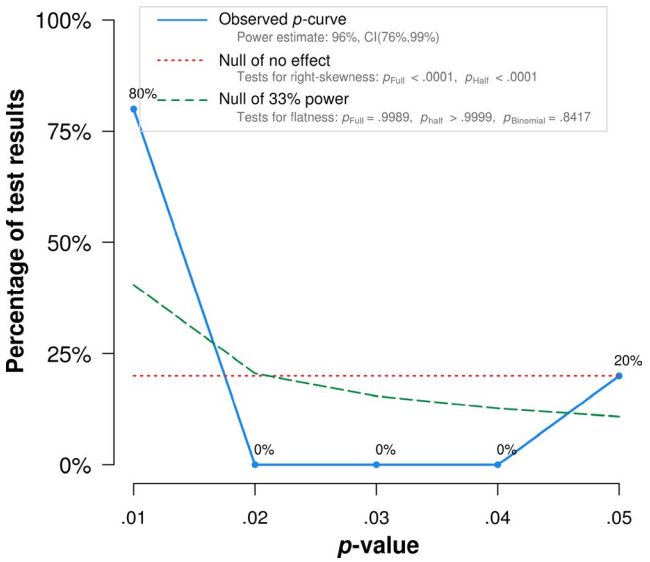

FIGURE 2 Small cross-modal numerosity discrimination p-curve. The blue line shows the distribution of *p*-values from the data. The analysis is run on *p*-values in a continuous manner. If the null hypothesis of zero effect is true, the distribution of *p*-values < .05 will be equal (red line). If the null hypothesis is true and the studies lack evidential value, the p-curve will be flatter than 33% (green line). If there is evidence of p-hacking, the p-curve will be left-skewed

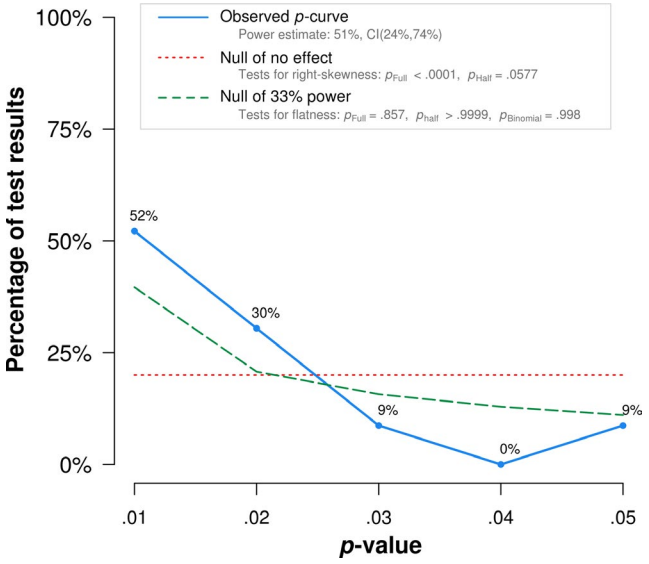

FIGURE 3 Large unimodal numerosity discrimination p-curve. The blue line shows the distribution of *p*-values from the data. The analysis is run on *p*-values in a continuous manner. If the null hypothesis of zero effect is true, the distribution of *p*-values < .05 will be equal (red line). If the null hypothesis is true and the studies lack evidential value, the p-curve will be flatter than 33% (green line). If there is evidence of p-hacking, the p-curve will be left-skewed

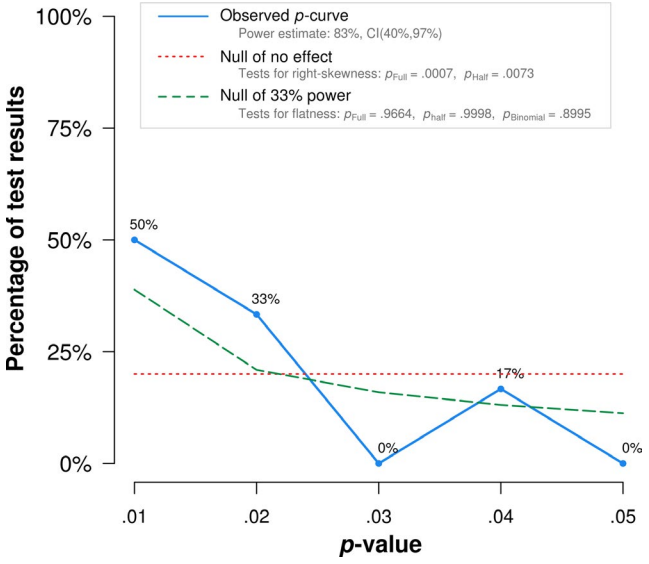

FIGURE 4 Large cross-modal numerosity discrimination p-curve. The blue line shows the distribution of *p*-values from the data. The analysis is run on *p*-values in a continuous manner. If the null hypothesis of zero effect is true, the distribution of *p*-values < .05 will be equal (red line). If the null hypothesis is true and the studies lack evidential value, the p-curve will be flatter than 33% (green line). If there is evidence of p-hacking, the p-curve will be left-skewed

most extreme *p*-values included from the full and the half p-curve analysis. In each cumulative p-curve analysis, the analysis stops when half of the *p*-values have been removed and, as such, each cumulative p-curve analysis may contain a different number of dropped *p*-values. This test allows for the examination of the stability of the p-curve to leaving out extreme *p*-values (both high and low).

3.2.1 | Small Numerosity discrimination (unimodal)

The evidential value demonstrated by the full p-curve and half p-curve for small unimodal numerosity discrimination remains unaffected when the most extreme *p*-values are dropped. The results of the cumulative p-curve analysis for small unimodal numerosity discrimination are displayed in Figure S1.

3.2.2 | Small Numerosity discrimination (cross-modal)

The evidential value of the small cross-modal numerosity discrimination p-curve also remains unaffected when the most extreme *p*-values are removed in the full and half p-curve conditions. This is shown in Figure S2.

3.2.3 | Large numerosity discrimination (unimodal)

The evidential value of the large unimodal numerosity discrimination full p-curve is unaffected by the removal of the most extreme *p*-values. Conversely, because the large unimodal numerosity discrimination p-curve only meets one adequacy condition (full p-curve and half p-curve significant at a level of $p < .1$), and the half p-curve is not significant at a level of $p < .05$, with the removal of the smallest *p*-values, the *p*-value of the half p-curve continues to increase. The results of the cumulative p-curve analysis for large unimodal numerosity discrimination are shown in Figure S3.

3.2.4 | Large numerosity discrimination (cross-modal)

Finally, the evidential value of the large cross-modal numerosity discrimination p-curve does not withstand the removal of the most extreme *p*-values for either the full p-curve or the half p-curve. The full p-curve stops demonstrating evidential value after the removal of the second smallest p-value, whereas the half p-curve stops demonstrating evidential value with the removal of the smallest p-value. This can be seen in Figure S4.

These analyses therefore suggest that evidential value in support of the small numerosity hypothesis appears to be more robust than the evidential value in support of the large numerosity hypothesis. As stated in the explanation of this analysis on the p-curve website (www.p-curve.com): *We should place more confidence in sets of studies whose overall evidential value survives the exclusion of the most extreme few results.*

3.3 | Robustness analyses

As described by Simonsohn et al. (2014a), robustness analyses involve re-running the p-curve analysis and replacing the chosen *p*-values with other dependent *p*-values measuring the same hypothesis of

interest from the experiment. Some studies report multiple *p*-values addressing the same hypothesis. However, in the p-curve analysis, only one dependent *p*-value per study can be included. If more than one *p*-value per study addressing a particular hypothesis exist, the analyses can be re-run with the alternate *p*-values and this serves to examine the robustness of the p-curve.

As such, a robustness analysis was done for each of the p-curves, the small unimodal and cross-modal numerosity discrimination and the large unimodal and cross-modal numerosity discrimination. A column in each of the p-curve disclosure tables displays the robustness results. When multiple *p*-values were available for the robustness analysis, the final *p*-value reported was used as described by Simonsohn et al. (2014a). In the small unimodal numerosity discrimination robustness analysis, three *p*-values were replaced by three *p*-values (in one study, the *p*-values for two ratios were collapsed into one *p*-value within the robustness calculation), and the robustness p-curve continued to meet both adequacy conditions and demonstrate evidential value. In the small cross-modal numerosity discrimination robustness analysis, three *p*-values were replaced by three *p*-values, and the robustness p-curve continued to meet both adequacy conditions and demonstrate evidential value. In the large unimodal numerosity discrimination robustness analysis, three *p*-values were replaced, and, the half p-curve remained not significant at a level of $p < .05$ and the full p-curve and the half p-curve continued to demonstrate significance at a level of $p < .1$. In the large cross-modal numerosity discrimination robustness analysis, three *p*-values were replaced by three *p*-values, and the robustness p-curve continued to meet both adequacy conditions and demonstrate evidential value.

Using the newly developed p-curves from the robustness analyses, cumulative analyses, in which extreme *p*-values are sequentially dropped (starting with the most extreme *p*-values), were again run. This process mirrored the cumulative analyses described above for the original p-curves, with the exception of the *p*-values used to create the p-curves. The cumulative analysis of the small unimodal numerosity robustness results demonstrate unchanged values when the smallest *p*-values are dropped, supporting the finding that evidential value exists. Like the cumulative analysis with the original large unimodal numerosity discrimination data, the cumulative analysis of the robustness analysis for large numerosity discrimination demonstrates that the half p-curve becomes non-significant as the lowest *p*-values are removed (>.05). In each of the cumulative analyses for the small and large cross-modal numerosity discrimination robustness analyses, the full and half p-curves become non-significant with the removal of the lowest *p*-values. These results suggest that we may be less confident in the robustness analyses of the cross-modal p-curves. Overall, the robustness analyses support the results obtained by the initial p-curve analyses for the hypotheses of both small and large numerosity discrimination.

4 | DISCUSSION

The infant numerosity discrimination literature has become foundational in providing support for the hypothesis that human infants possess an innate sense of numerosity (Feigenson et al., 2013). In order for the literature to be used effectively in shaping the theoretical underpinnings of empirical investigations into the development of the numerical processing system, it is imperative that the available evidence is assessed for evidential value in a method that considers publication bias and p-hacking. The p-curve analyses in this study provide an assessment of the evidential value of the numerosity discrimination literature upon which researchers have in part relied in establishing a basis for the numerosity system in human infants, while considering publication biases and the average power of the published studies. While other tasks, such as those assessing infants' arithmetic competence, may indicate successful numerosity discrimination, their inclusion in this type of analysis could have caused confusion when interpreting these results as processes beyond numerosity discrimination may be required to complete those tasks (e.g., computation of arithmetic operations). Additionally, Christodoulou et al. (2017) recommend further research to better understand the mechanisms involved in infants' ability to detect and preference for 'mathematically incorrect' stimuli in infant arithmetic tasks.

The p-curve analyses reported here demonstrate evidential value in the literature investigating infants' small and large unimodal and cross-modal numerosity discrimination. While each of small and large unimodal and cross-modal numerosity analyses demonstrate overall evidential value, a closer look at the data suggests nuanced results and differences in evidential value based on differences between the various analyses and supported by estimates of the underlying power. The power estimates vary greatly for the two unimodal hypotheses being examined, and it is of interest that the power for the large unimodal numerosity discrimination p-curve is lower than for the small unimodal numerosity discrimination p-curve. The confidence interval surrounding the power estimate is also wider for the large unimodal numerosity discrimination p-curve. Notably, the confidence interval range falls below 33% for the large but not small infant numerosity discrimination data. The p-curve for 33% power is used in p-curve analysis as the threshold to determine whether evidential value is inadequate or absent. Smaller, and less precise (wider CI's), power estimates are suggestive of more heterogeneity in the samples included in the p-curve analysis, resulting in less certainty about the power of the underlying effect.

Estimates of power for small and large cross-modal numerosity discrimination are higher than those for small and large unimodal numerosity discrimination. Despite higher estimates of underlying power in the cross-modal analyses, the cumulative analysis results of the robustness analyses reveal that with the removal of the lowest *p*-values, the cross-modal p-curves quickly become non-significant. This may lead to questions about the strength of the evidential value of the cross-modal numerosity discrimination literature. The same pattern is not observed for unimodal numerosity discrimination studies. The results of the main p-curve and the robustness results for unimodal numerosity discrimination are stable. In addition, the p-curve analyses demonstrate differences in power between the published small numerosity discrimination and the large numerosity discrimination data.

It is possible that stronger evidence for cross-modal discrimination is underpinned by the well-documented phenomenon of intersensory redundancy. According to the intersensory redundancy account, internal representations are strengthened through the presentation of redundant information across multiple senses (Jordan, Suanda, & Brannon, 2008). While this phenomenon does not apply to every cross-modal study (i.e., Kobayashi et al., 2005), it may account for some of the differences between unimodal and cross-modal discrimination. Also of importance, in differentiating these analyses, is the cognitive operation assessed through these tasks. In the unimodal tasks, infants are required to discriminate between quantities, whereas, in the cross-modal tasks, infants engage in matching corresponding amounts. This requires the discrimination of numerosities as well as the ability to relate (e.g., match) the numerosities to presentations across different modalities (Kobayashi et al., 2005). This task difference may explain differences in the amount of evidential value obtained. Future studies should contrast these different operations and how they affect evidence for infants' processing of both small and large numerosity.

The large unimodal numerosity discrimination p-curve met one of the two conditions that suggest evidential value in the data, the half p-curve and full p-curve demonstrate $p < .1$. In contrast, the p-curve for small unimodal numerosity discrimination met both conditions indicative of evidential value. The half p-curve demonstrates $p < .05$ and the half and full p-curve demonstrate $p < .1$. While meeting criterion for overall evidential value, the large unimodal numerosity discrimination analysis fails to meet the standards of evidential value based on the more robust test (the half p-curve), which leads to more ambiguous conclusions about the evidential value of the published large numerosity discrimination data (Simonsohn et al., 2015, p. 1150). The large unimodal numerosity discrimination p-curve also was not found to be robust to the removal of extreme *p*-values from the analysis.

One issue that may uniquely impact the interpretability of the small numerosity discrimination results is the way in which continuous perceptual variables, such as, surface area and contour, may confound numerosity discrimination in infants (Cantrell & Smith, 2013). We recognize the challenge that continuous variables create in interpreting whether the evidential value found in small numerosity discrimination is due to the ability to discriminate between numerosities or between changes in continuous variables. This question is not one we can answer within this p-curve analysis. The heterogeneity of the manipulations of continuous variables, within and across empirical studies, makes it difficult to assess the influence of continuous variables using meta-analytic techniques. Moving forward, replication studies, that systematically manipulate the continuous variables, could be of use to assess whether evidential value varies as a function of these experimental manipulations.

Taken together, the data then suggest that we can be relatively confident in the evidential value of the small unimodal numerosity data and the cross-modal data for both small and large numerosities, but less certain about the evidential value of the large unimodal numerosity data. It is important to note here that the results from this p-curve analysis do not imply that an approximate system for processing large unimodal numerosities does not exist in human infants. Rather the results of the present analysis suggest that more, and more robust, analysis is necessary to draw inferences about infants' large numerosity discrimination abilities unimodally.

With the available data, as it stands, no strong inferences in support of the hypothesis that infants can discriminate between large unimodal numerosities can be made. Therefore, adequately powered replication studies are required to clarify the evidential value of the data. Given the importance of the evidential value for current theories concerning the origins of numerical cognition, it is critical to conduct further studies to more unambiguously assess the evidential value of infant numerosity discrimination data. Efforts are being made to pursue this type of replication study in projects, such as, the Many Babies Project (Frank, 2017; http://babieslearninglanguage.blogspot.ca/2015/12/the-manybabies-project.html). Such replication efforts should also take into account recommendations put forward for improving the practices of infant research (see Eason, Hamlin, & Sommerville, 2017).

One may be tempted here to speculate on the reasons as to why, as per the p-curve analyses, there is greater evidential value for the hypothesis that infants can discriminate between small numerosities than there is for the hypothesis that infants can discriminate between large numerosities. One approach may be to look for moderators of the strength of the effect such as differences between studies in the way that the stimuli were constructed and presented or variability between studies regarding whether infants habituated to the stimuli in the first place. However, in view of the low power and the large heterogeneity of the statistical power in the published studies, we take the view that any consideration of such moderators must await the results of adequately powered replication studies to assess evidential value and consequently whether the null hypothesis can be rejected or not.

Once such studies have been conducted to more comprehensively quantify the evidence for infant numerosity discrimination abilities, particularly large unimodal numerosity discrimination, theories of the origins of numerical abilities will be able to be grounded in evidence that is stronger and more reliable than is currently the case. Until then, it is important that the present p-curve analyses are updated as new evidence is published, that similar analyses are performed investigating the evidential value of other early numerical competencies, also foundational to theories of numerical abilities, and, perhaps most importantly, that researchers in the field take the present state of the evidential value of hypotheses concerning infant numerosity discrimination data into account when formulating theories regarding the ontogenetic foundations of numerical competencies.

DATA AVAILABILITY

The data that support the findings of this study are openly available in The Open Science Framework: Small Numerosity

Discrimination (Unimodal) https://osf.io/wrv7j/ (https://doi.org/10.17605/OSF.IO/WRV7J; Smyth & Ansari, 2019a), Small Numerosity Discrimination (Cross-Modal) https://osf.io/ae57r/ (https://doi.org/10.17605/OSF.IO/AE57R; Smyth & Ansari, 2019b), Large Numerosity Discrimination (Unimodal) https://osf.io/hgfym/ (https://doi.org/10.17605/OSF.IO/HGFYM; Smyth & Ansari, 2019c), and Large Numerosity Discrimination (Cross-Modal) https://osf.io/cgh98/ (https://doi.org/10.17605/OSF.IO/CGH98; Smyth & Ansari, 2019d).

ORCID

Rachael E. Smyth https://orcid.org/0000-0001-6673-1824

ENDNOTE

[1] There are not, to our knowledge, a specific minimal number of studies required for this type of analysis to be meaningful. p-curves can be generated for single papers to assess the evidential value of the results reported. A result of evidential value demonstrates that the *p*-values included in that analysis are sufficiently right-skewed to exclude the possibility that the null hypothesis is true. A p-curve analysis with more *p*-values may provide a less conclusive outcome, such as, a lack of evidential value and a failure to meet inadequacy criteria. In these instances, more studies/replications may be required to determine whether the effect occurs with evidential value or whether previous findings may be the result of false positives under the null. Figure 6 in the Simonsohn et al. (2014a) paper, p-curve: a key to the file drawer depicts the frequency with which p-curves demonstrate evidential value when a true effect does not exist (false positive) and the frequency with which p-curves fail to demonstrate evidential value when a true effect does exist (false negative) based on the number of *p*-values used in the p-curve. Based on the number of *p*-values included in each of our analyses (smallest $n = 5$), if studies were powered at 80%, the p-curve should be able to detect evidential value of a true effect at least 92% of the time. If, for example, studies were powered at 50%, seven *p*-values in a p-curve would be able to detect a true effect 61% of the time.

REFERENCES

Ansari, D., Lyons, I. M., Van Eimeren, L., & Xu, F. (2007). Linking visual attention and number processing in the brain: The role of the temporo-parietal junction in small and large symbolic and nonsymbolic number comparison. *Journal of Cognitive Neuroscience*, *19*, 1845–1853. https://doi.org/10.1162/jocn.2007.19.11.1845

Antell, S. E., & Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child Development*, *54*, 695–701. https://doi.org/10.2307/1130057

Brannon, E. M., & Merritt, D. (2011). Evolutionary foundations of the approximate number system. In S. Dehaene & E. M. Brannon (Eds.), *Space, time, and number in the brain: Searching for the foundations of mathematical thought* (pp. 207–224). London, UK: Academic Press (Elsevier).

Brez, C. C. (2009). *Infant number perception: A developmental approach*. Retrieved from http://hdl.handle.net/2152/7551

Cantlon, J. F. (2012). Math, monkeys, and the developing brain. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(Suppl. 1), 10725–10732. https://doi.org/10.1073/pnas.1201893109

Cantrell, L., & Smith, L. B. (2013). Open questions and a proposal: A critical review of the evidence on infant numerical abilities. *Cognition*, *128*, 331–352. https://doi.org/10.1016/j.cognition.2013.04.008

Christodoulou, J., Lac, A., & Moore, D. S. (2017). Babies and math: A meta-analysis of infants' simple arithmetic competence. *Developmental Psychology*, *53*(8), 1405–1417.

Clearfield, M. W. (2004). Infants' enumeration of dynamic displays. *Cognitive Development*, *19*, 309–324. https://doi.org/10.1016/j.cogdev.2004.03.003

Eason, A. E., Hamlin, J. K., & Sommerville, J. A. (2017). A survey of common practices in infancy research: Description of policies, consistency across and within labs, and suggestions for improvements. *Infancy*, *22*, 470–491. https://doi.org/10.1111/infa.12183

Feigenson, L. (2011). Predicting sights from sounds: 6-month-olds' intermodal numerical abilities. *Journal of Experimental Child Psychology*, *110*, 347–361. https://doi.org/10.1016/j.jecp.2011.04.004

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

Feigenson, L., Carey, S., & Spelke, E. S. (2002). Infants' discrimination of number vs. continuous extent. *Cognitive Psychology*, *44*(1), 33–66. https://doi.org/10.1006/cogp.2001.0760

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

Feigenson, L., Libertus, M. E., & Halberda, J. (2013). Links between the intuitive sense of number and formal mathematics ability. *Child Development Perspectives*, *7*(2), 74–79. https://doi.org/10.1111/cdep.12019

Frank, M. C. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*, 421-435. https://doi.org/10.1111/infa.12182

Hyde, D. C. (2011). Two systems of non-symbolic numerical cognition. *Frontiers in Human Neuroscience*, *5*, 1–8. https://doi.org/10.3389/fnhum.2011.00150

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701. https://doi.org/10.1371/journal.pmed.0020124

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

Jordan, K. E., & Brannon, E. M. (2006). The multisensory representation of number in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3486–3489. https://doi.org/10.1073/pnas.0508107103

Jordan, K. E., Suanda, S. H., & Brannon, E. M. (2008). Intersensory redundancy accelerates preverbal numerical competence. *Cognition*, *108*, 210–221. https://doi.org/10.1016/j.cognition.2007.12.001

Kobayashi, T., Hiraki, K., & Hasegawa, T. (2005). Auditory-visual intermodal matching of small numerosities in 6-month-old infants. *Developmental Science*, *8*, 409–419. https://doi.org/10.1111/j.1467-7687.2005.00429.x

Lakens, D. (2014). Professors are not elderly: Evaluating the evidential value of two social priming effects through p-curve analyses. *SSRN Electronic Journal*, 1–13. https://doi.org/10.2139/ssrn.2381936

Libertus, M. E., & Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Developmental Science*, *13*, 900–906. https://doi.org/10.1111/j.1467-7687.2009.00948.x

Libertus, M. E., Starr, A., & Brannon, E. M. (2014). Number trumps area for 7-month-old infants. *Developmental Psychology*, *50*, 108–112. https://doi.org/10.1037/a0032986

Lipton, J., & Spelke, E. S. (2003). Origins of number sense: Large number discrimination in human infants. *Psychological Science*, *14*, 396–401. https://doi.org/10.1111/1467-9280.01453

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, *5*, 271–290. https://doi.org/10.1207/s15327078in0503_2

Mack, W. (2006). Numerosity discrimination: Infants discriminate small from large numerosities. *European Journal of Developmental Psychology*, *3*(1), 31–47. https://doi.org/10.1080/17405620500347695

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, *15*, 776–781. https://doi.org/10.1111/j.0956-7976.2004.00755.x

Moore, D., Benenson, J., Reznick, J. S., Peterson, M., & Kagan, J. (1987). Effect of auditory numerical information on infants' looking behavior: Contradictory evidence. *Developmental Psychology*, *23*, 665–670. https://doi.org/10.1037/0012-1649.23.5.665

Nieder, A. (2016). The neural code for number. *Nature Reviews Neuroscience*, *17*, 366–382. https://doi.org/10.1016/B978-0-12-385948-8.00008-6

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, *32*, 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550

Núñez, R. E. (2017). Is there really an evolved capacity for number? *Trends in Cognitive Sciences*, *21*, 409–424. https://doi.org/10.1016/j.tics.2017.03.005

Piazza, M., & Eger, E. (2016). Neural foundations and functional specificity of number representations. *Neuropsychologia*, *83*, 257–273. https://doi.org/10.1016/j.neuropsychologia.2015.09.025

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-curve and effect size: Correcting for publication bias using only significant results. *Psychological Science*, *9*, 666–681. https://doi.org/10.1177/1745691614553988

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*, 1146–1152. https://doi.org/10.1037/xge0000104

Smyth, R. E., & Ansari, D. (2019a). Small numerosity discrimination (unimodal); OSF; version 2. https://doi.org/10.17605/OSF.IO/WRV7J

Smyth, R. E., & Ansari, D. (2019b). Small numerosity discrimination (cross-modal); OSF; version 1. https://doi.org/10.17605/OSF.IO/AE57R

Smyth, R. E., & Ansari, D. (2019c). Large numerosity discrimination (unimodal); OSF; version 2. https://doi.org/10.17605/OSF.IO/HGFYM

Smyth, R. E., & Ansari, D. (2019d). Large numerosity discrimination (cross-modal); OSF; version 1. https://doi.org/10.17605/OSF.IO/CGH98

Starkey, P., & Cooper, R. (1980). Perception of numbers by human infants. *Science*, *210*, 1033–1035.

Starkey, P., Spelke, E. S., & Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. *Science*, *222*, 179–181.

vanMarle, K., & Wynn, K. (2009). Infants' auditory enumeration: Evidence for analog magnitudes in the small number range. *Cognition*, *111*, 302–316. https://doi.org/10.1016/j.cognition.2009.01.011

Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition*, *89*(1), 15–25. https://doi.org/10.1016/S0010-0277(03)00050-7

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Smyth RE, Ansari D. Do infants have a sense of numerosity? A p-curve analysis of infant numerosity discrimination studies. *Dev Sci*. 2020;23:e12897. https://doi.org/10.1111/desc.12897