Cognitive Development 52 (2019) 100817

Contents lists available at ScienceDirect

Cognitive Development

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

Certainty in numerical judgments develops independently of the approximate number system

Carolyn Baer*, Darko Odic

University of British Columbia, Canada

ARTICLE INFO

Keywords:
Certainty
Confidence
Approximate number system
Discrimination

ABSTRACT

Recent work has shown that the precision with which children reason about their ANS certainty improves with age: when making simple number discrimination decisions, like deciding whether there are more blue or yellow dots on the screen, older children are better able to differentiate trials that they answered correctly vs. incorrectly. Here, in two experiments, we examine whether the age-related improvement in ANS certainty is accounted for by children's: (1) increasing ability to properly "calibrate" their certainty judgements (i.e., a reduction in over-confidence with age); (2) improving precision of the ANS representations themselves; and/or (3) the improvement of children's ability to represent and reason about certainty in general. By testing children in a child-friendly "relative" certainty task, we find that 3–7 year-olds' (N = 161) certainty in their ANS decisions develops independently of both ANS acuity and calibration abilities. These results hold even when non-numeric perceptual features, such as the density and cumulative area, are controlled for. We discuss these results in a broader context of children's general ability to reason about certainty and confidence.

1. Introduction

Imagine coming back from a concert and being asked how many people you think attended the show. Even though you (hopefully) did not spend time counting each individual person there, research over the past twenty years has shown that you could, without difficulty, roughly estimate the number of people you saw by using your Approximate Number System (ANS; Dehaene, 2011; Halberda, Ly, Wilmer, Naiman, & Germine, 2012; Odic & Starr, 2018; c.f. Gebuis & Reynvoet, 2012; Szűcs, Nobes, Devine, Gabriel, & Gebuis, 2013). The ANS is the theorized evolutionarily-adapted system for representing numerical information that guides our earliest intuitions about number. It is present in newborn infants (Izard, Sann, Spelke, & Streri, 2009), preschoolers (Halberda & Feigenson, 2008), and many non-human animals (for review, see Vallortigara, 2017). The key signature of the ANS is that it represents number imprecisely, following Weber's law that discriminability is linked to the ratio between numbers (Weber, 1978). That is, given a large ratio between two numbers (e.g., groups of 10 vs. 20 dots on a screen), we can easily tell their difference; but, given a smaller ratio (e.g., 10 vs. 11 dots), the underlying noise of the ANS representations is too high to reliably tell which group has more dots. Over the course of development, the internal precision of the ANS slowly improves – peaking sometime between late adolescence and adulthood (Halberda et al., 2012; Odic, 2018) – allowing us to make increasingly accurate intuitive number judgments, even in the absence of counting or language.

Recent theoretical and empirical work has shown that the ANS provides us with both an approximate sense of number *and* a sense

* Corresponding author at: Department of Psychology, University of British Columbia 2136 West Mall Vancouver, British Columbia V6T 1Z4, Canada.
E-mail address: cebaer@psych.ubc.ca (C. Baer).

https://doi.org/10.1016/j.cogdev.2019.100817
Received 28 November 2018; Received in revised form 23 August 2019; Accepted 27 August 2019

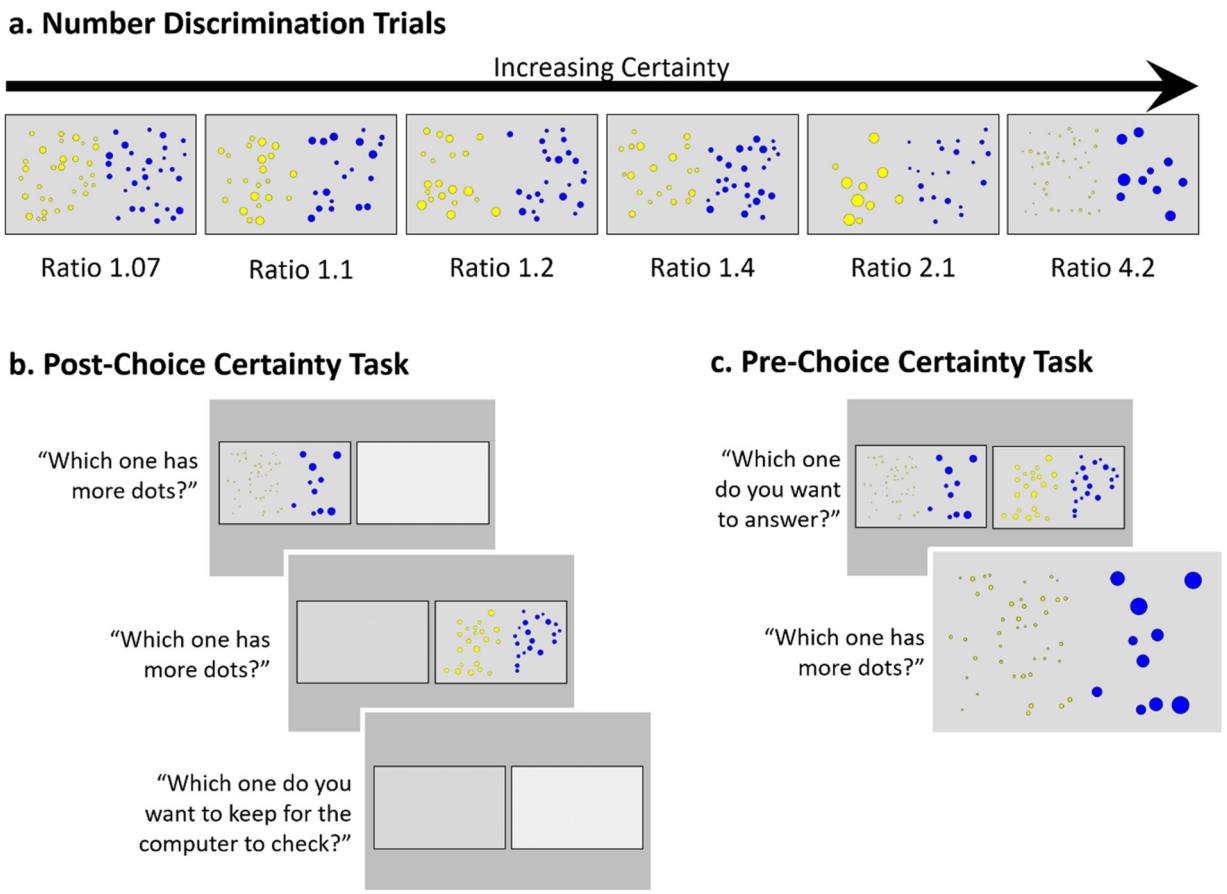

Fig. 1. Sample stimuli used in the study. Section a depicts sample number discrimination trials in which children have to indicate which color has more dots. Section b depicts the Post-Choice Certainty Task, in which children first answer the question on the left, then the question on the right, then are asked to select the answer they were most confident in. Section c depicts the Pre-Choice Certainty Task, in which children first answer the certainty question by selecting the trial they most expect to get correct, then answer only that question.

of our certainty in that estimate. For example, if you were asked to estimate the number of words on this page, your ANS would provide you with both the most likely number but also a sense of how confident you should be in that value (Halberda & Odic, 2014; Vo, Li, Kornell, Pouget, & Cantlon, 2014). Young children can also reason about their certainty in simple ANS decisions: after completing a number discrimination trial (deciding whether there are more dots on the left or right side in Fig. 1) 5–8 year-old children can also indicate whether they believe that they answered the trial correctly or incorrectly by choosing a value on a 2-point scale (Vo et al., 2014; see also Baer, Gill, & Odic, 2018). This work is consistent with broader work demonstrating that young children and toddlers can monitor and use or report their uncertainty in a variety of cognitive and perceptual tasks, including when identifying objects (Lyons & Ghetti, 2011), remembering novel names (Lipowski, Merriman, & Dunlosky, 2013), or deciding whether they are confident enough to walk down a narrow ramp (Tamis-LeMonda et al., 2008).

But, while children have an early ability to reason about certainty in their ANS decisions, they are far from perfect at it. For example, children are not always well "calibrated" in their certainty ratings and are often over-confident in their estimates of their knowledge or their accuracy on numerical and non-numerical tasks (Lipowski et al., 2013; van Loon, de Bruin, Leppink, & Roebers, 2017; Vo et al., 2014). Furthermore, children's "sensitivity" to certainty – their ability to tell apart increasingly similar states of certainty (e.g., both the difference between "sure" vs. "unsure", as well as the more nuanced difference between "sure" vs. "somewhat sure"), sometimes called "resolution" – also develops and becomes finer tuned with age for both their ANS and other domains where certainty has been tested (Vo et al., 2014, though see Salles, Ais, Semelman, Sigman, & Calero, 2016). As a result, while older children are generally quite good at differentiating between the trials that they answered correctly vs. incorrectly, reflecting good sensitivity to certainty, younger children are significantly poorer.

If children are becoming more sensitive to their certainty in ANS decisions, which factors predict this developmental trajectory? In other words, *why* does children's sensitivity to ANS certainty improve with age? As we explain in detail below, we consider and test three possible explanations for this developmental change: (1) that children's sensitivity to certainty in their ANS decisions only *appears* to improve because of improvements in calibration (e.g., a reduction in being over-confident; Salles et al., 2016); (2) that children's improving sense of certainty is accounted for by the improvements in the ANS representations themselves – i.e., the

reduction in the perceptual noise from which certainty may be extracted (e.g., Maniscalco & Lau, 2012; Maniscalco & Lau, 2014); and (3) that children's improving sense of certainty in their ANS decisions is accounted for by more general improvement in reasoning and representing certainty independent of the ANS itself (e.g., Baer et al., 2018).

Under the first hypothesis, children's sensitivity to certainty may reach adult-like levels at a very young age, but this early competency may be overshadowed by children's poor calibration abilities. In typical paradigms measuring certainty abilities (e.g., Lyons & Ghetti, 2011; Salles et al., 2016; Vo et al., 2014), children are first asked to make a simple decision, such as guessing whether there are more blue or yellow dots (as in Fig. 1), followed by a second question asking them to rank their certainty on a scale (e.g., "sure" vs. "not sure"). While tasks such as these are intuitively and methodologically straightforward, decades of research in the study of certainty more broadly have shown that they tap into both individual differences in certainty sensitivity *and* individual differences in where participants set their internal criterion for what counts as high vs. low certainty (Barthelmé & Mamassian, 2009; Butterfield, Nelson, & Peck, 1988; Lipowski et al., 2013; Nelson, 1984; Salles et al., 2016). As a result, children's improving performance on these tasks could either be evidence of improving ability to differentiate correct from incorrect trials (i.e., sensitivity), or of better and less overly optimistic criterion-setting (i.e., calibration). Consistent with this, Salles et al. (2016) show that when signal-detection approaches are used to disentangle sensitivity (i.e., *d'*) from criterion-setting, children show adult-like certainty sensitivity by age 6 in surface area discrimination tasks, but continue developing their criterion-setting well beyond this age. Under this hypothesis, what appears to be the development of sensitivity may actually be an improvement in calibration, and we should, therefore, find little-to-no development of sensitivity in ANS certainty when controlling for young children's poor certainty calibration.

Under the second hypothesis, children's developing sensitivity in ANS certainty may simply be a by-product of the improving ANS representations themselves. By analogy, consider a task in which you must choose the darker of two shades of grey and then subsequently report your certainty in that decision. In other words, you must make two decisions: the underlying perceptual decision of selecting the darker shade (sometimes termed a "Type 1 decision"), and a certainty report (sometimes termed a "Type 2 decision"; Galvin, Podd, Drga, & Whitmore, 2003; Maniscalco & Lau, 2012; Maniscalco & Lau, 2014). Crucially, these decisions necessarily interact: if your perceptual system for seeing brightness is very imprecise (e.g., you see all greys as either black or white), your Type 1 decision will always be either impossible (as all nearby shades of gray will seem identical) or trivially easy (e.g., when the two shades are extremely different). Thus, with a highly noisy perceptual system, your Type 2 certainty decision would also *appear* imprecise because it also only has two states: impossible or trivially easy. Developing perceptual abilities themselves might, therefore, be partially or completely responsible for improving sensitivity to certainty (Maniscalco & Lau, 2014; Maniscalco & Lau, 2012; Pouget, Drugowitsch, & Kepecs, 2016). For example, cumulative area and ANS representations (like those used by O'Leary & Sloutsky, 2017; Salles et al., 2016; Vo et al., 2014) heavily develop and become increasingly precise between birth and early adolescence (Halberda & Feigenson, 2008; Halberda et al., 2012; Odic, 2018; Piazza, De Feo, Panzeri, & Dehaene, 2018), opening the possibility that children's improving certainty sensitivity in these dimensions could itself be entirely due to the reduction of noise in these underlying perceptual representations. Therefore, under this hypothesis, we should find little-to-no unique development in sensitivity to certainty when controlling for individual and developmental differences in children's underlying perceptual representations.

Finally, under the third hypothesis, children's sensitivity to ANS certainty might improve through more general mechanisms that actively extract, represent, and use certainty information and themselves improve over time. Two recent lines of work suggest that representations of certainty may tap into more general factors that go beyond the Type 1 decision observers are making. First, work with adult participants has shown that certainty decisions can be easily compared across otherwise independent perceptual tasks, including across independent modalities such as vision and audition (De Gardelle, Le Corre, & Mamassian, 2016; De Gardelle & Mamassian, 2014), suggesting that certainty may use or even be represented on a domain-general scale. Recently, this work has been extended developmentally, finding that while number, area, and emotion perception are representationally independent (i.e., how well a child discriminates number does not predict how well they discriminate emotional expressions), their certainty judgements are tightly correlated and constitute a single factor across these three dimensions (Baer et al., 2018). Therefore, children's increasing sensitivity in their certainty decisions may be the by-product of a developing domain-general certainty scale, and not specifically tied to the developing ANS itself. Additionally, recent computational models of certainty in adults have suggested that certainty may in part depend on the momentary noise in perceptual information, but also on a host of other performance factors that observers combine to determine how sure they are of a momentary decision, such as their general confidence in the task at hand, their prior history of trials, how much attention they believe they were dedicating to the current trial, etc. (Martí, Mollica, Piantadosi, & Kidd, 2018; Pouget et al., 2016). Children may, therefore, become more sensitive at judging their ANS certainty because they are better able to combine a variety of relevant cues when deciding whether they answered a particular trial correctly or incorrectly. Under either of these views, we should find that children's sensitivity to certainty should continue to develop even when controlling for individual and developmental differences in calibration and the underlying noise in the ANS, as children's certainty decisions themselves should be a product of factors that are at least partly independent of the ANS itself.

To tease these possibilities apart, we need a method that both controls for children's criterion-setting and that can allow us to measure the precision of the ANS independently from certainty decisions. To accomplish this, we elected to use a "relative" certainty task, also sometimes termed a Forced-Choice Certainty task (Barthelmé & Mamassian, 2009; Mamassian, 2016), which is popular in the adult literature on certainty perception as it directly measures certainty sensitivity independently of potential response biases. In relative certainty tasks, participants are asked to first answer two Type-1 questions (e.g., two perceptual discrimination trials, such as deciding whether there are more blue or yellow dots in Fig. 1, from which we can measure ANS precision itself), and are then asked to report which question they are *more* certain of getting correct (Barthelmé & Mamassian, 2009; Butterfield et al., 1988; Lipowski et al., 2013). Following the principles of Signal Detection Theory (SDT; Green & Swets, 1966), relative tasks such as these allow researchers to measure sensitivity to certainty independent of criterion-setting. Because the observer simply compares two internal states and

Table 1
Sample sizes, means, tests against chance, and model fit estimates for the Number task, the Post-Choice version, and the Pre-Choice version in Experiment 1.

Age	*N*	% Correct (*SD*)	*t*	*p*	*d*
Number Discrimination Task					
Overall	98	79.52 (10.83)	26.99	< .001	2.73
3	13	68.72 (9.26)	7.29	< .001	2.02
4	15	68.44 (13.41)	5.33	< .001	1.38
5	20	79.67 (7.90)	16.79	< .001	3.75
6	22	83.86 (5.45)	29.14	< .001	6.21
7	28	86.96 (5.05)	38.71	< .001	7.31
Post-Choice Certainty Task					
Overall	98	60.48 (16.18)	6.41	< .001	0.65
3	13	51.54 (9.87)	0.56	.585	0.15
4	15	53.33 (10.69)	1.21	.247	0.31
5	20	58.33 (15.20)	2.45	.024	0.55
6	22	63.18 (18.50)	3.34	.003	0.71
7	28	67.86 (16.64)	5.68	< .001	1.07
Pre-Choice Certainty Task					
Overall	99	60.74 (16.10)	6.64	< .001	0.67
3	13	49.49 (10.79)	−0.17	.867	−0.05
4	14	50.24 (11.58)	0.08	.940	0.02
5	20	60.17 (13.00)	3.50	.002	0.78
6	22	63.03 (12.64)	4.84	< .001	1.03
7	30	69.22 (18.77)	5.61	< .001	1.02

decides which one they are *more* sure of, they are not forced to set a criterion value for what counts as low vs. high certainty at all, allowing us to measure certainty sensitivity directly and independently of criterion-setting. Thus, if children's sensitivity to ANS certainty peaks independent of poor criterion-setting, we should find that children do not show improvements in the relative certainty task past preschool (Salles et al., 2016). In contrast, if we find continued development in sensitivity in the relative certainty task, we would have evidence that poor criterion-setting is not solely responsible for these changes.

2. Experiment 1

2.1. Methods

2.1.1. Participants

We opportunistically tested a total of 100 children (M = 5; 11, range = 3; 2–8; 0 [years; months], 56 girls), an arbitrary sample size chosen a priori (see Table 1 for distributions by age). All children were tested in a quiet space in their schools or daycares in [location blinded for review]. No additional demographic information was collected, though most children were middle- to upper-middle class and largely from [ethnicity blinded for review] backgrounds.

2.1.2. Materials and procedures

Tasks were presented on an 11.3" Apple Air laptop computer using Psychtoolbox-3 (Brainard, 1997). Children could respond by verbally indicating their choice, or by pointing to a side of the screen. The experimenter pushed all buttons to reduce the influence of memory and motor development on the results.

Stimuli throughout the experiment consisted of trials from a number discrimination task used widely in the literature on the ANS (Halberda, Mazzocco, & Feigenson, 2008; Odic, 2018; Odic & Starr, 2018). In each trial, there are two spatially separated groups of dots that differ in number, and children are asked to determine (without counting) whether there are more blue or yellow dots on the screen (see Fig. 1). The size of the dots within each screenshot and across screenshots was varied to control for the cumulative area of the dots. Children who attempted to count the dots were reminded of the no counting rule, and the experimenter covered the dots with her hand if the child continued. We manipulated children's probability of getting a trial correct by adjusting the ratio between the two sets of dots (see Halberda & Feigenson, 2008; O'Leary & Sloutsky, 2017; Vo et al., 2014). For instance, the last image in Fig. 1a depicts a ratio of 4.2 (42 yellow dots and 10 blue dots), which elicits a high degree of certainty, and which most children in this age range would get correct (Odic, 2018). In contrast, the first image in Fig. 1a depicts a ratio of 1.07, which elicits a much lower degree of certainty, and which very few children in this age range get correct above chance. Each trial varied continuously in ratio from 1.05 to 5.0, binned into 6 groups: 1.07, 1.10, 1.23, 1.44, 1.92, and 4.17.

Before starting the study, children completed 9 practice number discrimination trials presented on flashcards to teach them how to complete the number discrimination task. Practice trial ratios ranged from 1.33 to 3, and children were told whether their answers were correct or not. Then, children were told they would play the game 'for real' on the computer, and they needed to get a lot of questions right to win.

To assess children's certainty sensitivity, we designed two versions of the relative certainty task (described in detail below). In one

version, modelled directly off the Forced Choice tasks used with adults (e.g., Barthelmé & Mamassian, 2009; De Gardelle & Mamassian, 2014), children first answered two number discrimination questions, then indicated which answer had higher certainty. In the second version, children were shown two trials *simultaneously* and then selected the one they were most certain of to answer (for a similar approach, see Barthelmé & Mamassian, 2009, Study 1; Baer et al., 2018). We will refer to the first version as the "Post-Choice" Certainty task because the certainty judgment is made *after* the perceptual decisions, and the second version as the "Pre-Choice" Certainty task because the certainty judgment is made *before* the perceptual decision.

Each version of the relative certainty task has its own strengths and limitations. The Post-Choice task allows us to simultaneously collect certainty and perceptual judgments for all trials, while the Pre-Choice task does not, because children only answer the one question they indicate as high certainty. However, the Post-Choice version potentially places additional cognitive and motivational demands on children that the Pre-Choice version does not. Completing the Post-Choice task requires that children hold in memory their two states of certainty from the preceding perceptual decisions, overcome cognitive fatigue to report on their certainty after answering both perceptual decisions, and stay motivated through the task without evaluative feedback (feedback about the accuracy of their perceptual judgments would eliminate the need for children to consult their certainty - they could simply choose the question they received positive feedback on). Despite these differences, we hypothesized that both tasks would measure the same underlying abilities. We therefore ran both versions on all children, counterbalancing order across participants. All but 3 children completed both versions: children who only completed one version were retained for analyses of that task, but were removed for comparisons between the two versions. Children were permitted to take a short stretching break in between tasks to reduce boredom.

2.1.2.1. Post-choice certainty task. In this task, children were shown two gray occluders – one on the left side of the screen and one on the right (Fig. 1b). When the child was ready, the experimenter pushed a button to reveal a picture of blue and yellow dots behind the left occluder and the child was asked whether there are more blue or yellow dots. Children could either point or verbally indicate which set had more dots, after which point the experimenter would push a button and the trial would get re-covered by the occluder. Children were given as long as they needed to answer which color had more dots, but they were told not to count and were prevented from counting if they ignored this rule. After the child answered the first trial, the experimenter would push a button to reveal the second picture of blue and yellow dots, and the child would again answer which side had more. No feedback was given about the accuracy of the answer, as this could have changed children's certainty judgments. Instead, the experimenter occasionally gave neutral encouraging affirmations ("Okay!", "Alright!") to keep children engaged, ensuring to always provide equivalent feedback for both left and right answers.

After answering both questions, the experimenter asked the child "Which one do you want to keep for the computer to check? Which one are you more sure of?". Variations of these questions have been successfully used to elicit certainty judgments in children as young as 3 years (Hembacher & Ghetti, 2014; Vo et al., 2014). As in Barthelmé and Mamassian (2009), children were not able to see the questions during this phase (though they could still see the occluders) and had to rely on the memory of their certainty. The experimenter did not provide any feedback about whether their selected question was correct or not, as this feedback might also have been interpreted as indicating that their *certainty* choice was correct or not (see Smith, Beran, Couchman, & Coutinho, 2008).

Critically, the trials were paired such that one always displayed a larger (i.e., higher certainty) ratio than the other. We expected, based on other work with this paradigm, that children would choose the answer they felt was more certain (Baer et al., 2018). To assess individual differences in sensitivity to certainty, we varied the relative difference between the ratios of the two presented trials, which we quantified with a "metaratio": the larger ratio divided by the smaller one (e.g., metaratio 4.0 could be made with ratio 4.2 and ratio 1.05). On each trial, children were presented with one of five metaratios: 4.0, 3.0, 1.5, 1.25, or 1.1. Each metaratio was presented 6 times, yielding a total of 30 trials. All 60 number discrimination trials used to make the certainty trials were unique. Note that rather than using a division of ratios, we could have instead calculated the difference of ratios; both ratio and difference approaches have previously been used in the literature (e.g., De Gardelle, Le Corre, & Mamassian, 2016), and our choice of using division does not impact any of our results and was chosen to remain consistent with previous reports (Baer et al., 2018).

Our two primary dependent variables of interest in this task were each child's accuracy in identifying which set had more dots on each of the 60 trials (i.e., number discrimination accuracy) and the child's choice of which trial to keep on the certainty questions – i.e., which trial they were more certain of.

This task took, on average, 5.6 min for children to complete.

2.1.2.2. Pre-choice certainty task. The stimuli for this version were identical to the Post-Choice version: the identical 60 number discrimination trials were used in exactly the same pairings as in the Post-Choice version to limit the differences between the tasks. However, in the Pre-Choice version, both number discrimination trials were visible side-by-side on the screen at the beginning of each trial (Fig. 1c). Rather than answering each question and then retrospectively evaluating their certainty, children were instead asked "Which one do you want to do?" (this prompt has successfully elicited certainty judgements from children in previous work as children seek to maximize their success; Baer et al., 2018). Their selected question would then expand to fit the whole screen, hiding the non-chosen option, and they indicated the side with more dots. In other words, children evaluated their certainty prospectively and chose a trial to complete based on their perceived higher certainty. Children were given as long as they needed to answer both the certainty and number discrimination questions, though they were discouraged from counting in the same way as in the Post-Choice version. To maintain engagement, children were given feedback on whether they got the answer correct in the zoomed-in number discrimination (e.g., "That's right!" or "Oh, that's not right!"), as there was no way for this feedback to be misinterpreted as feedback about their certainty choice.

The primary dependent variable in this task was the trial that children choose to attempt – i.e., the one they were more certain in.

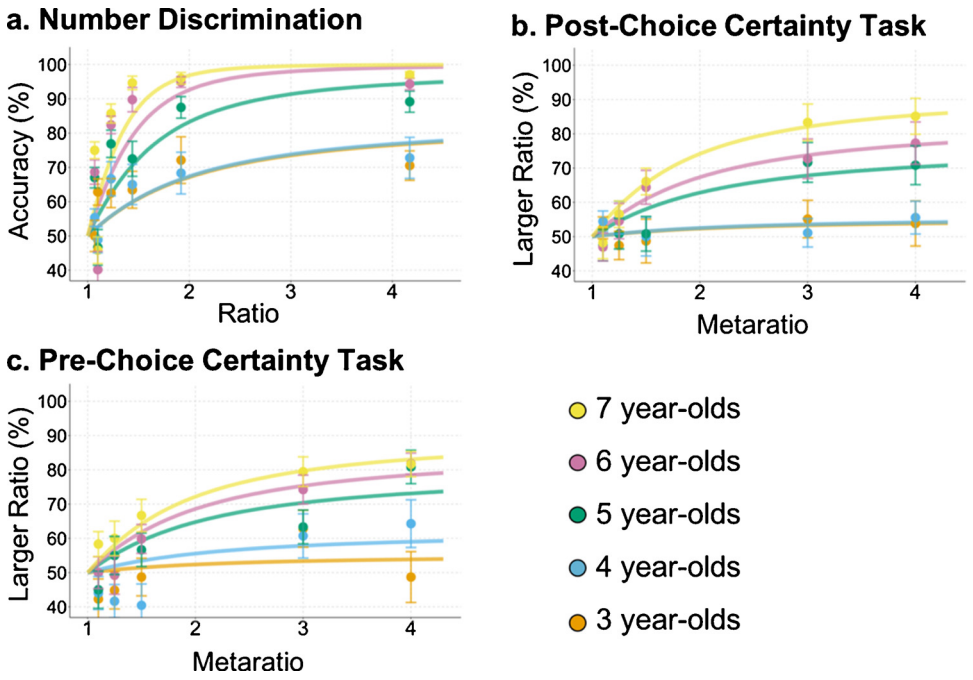

Fig. 2. Accuracy at each ratio on the number discrimination trials, and at each metaratio certainty trials in the Post-Choice and Pre-Choice Certainty Tasks in Experiment 1. Error bars represent 1 SE, and curves are estimated using a standard psychophysical model (see Odic, 2018).

This task took, on average, 3.6 min for children to complete.

2.2. Results

We found no effects of gender in our analyses, so all results reported hereafter collapse across gender. Children were generally more accurate on the Pre-Choice Certainty Task if they completed it first, likely because the longer Post-Choice task was more fatiguing, $F(1, 92) = 4.73, p = .032, \eta_p^2 = .05$. We report the remainder of the results combined across orders, as no results change if we include it. All ANOVAs are Greenhouse-Geisser corrected if sphericity is violated.

2.2.1. Number discrimination

Children's average accuracy on the number discrimination trials within the Post-Choice task was 80% ($SD = 11\%$), which was significantly higher than chance, $t(97) = 26.99, p < .001, d = 2.73$. This level of performance is consistent with previously reported ANS performance in this age range (Odic, 2018). Consistent with the classic ratio-dependent signature of the ANS, children were more accurate, $F(3.33, 322.61) = 83.87, p < .001, \eta_p^2 = .46$; see Fig. 2, and faster, $F(3.91, 379.21) = 12.38, p < .001, \eta_p^2 = .11$, on larger ratios compared to smaller ones. Finally, there was a significant correlation between age and number discrimination accuracy, $r(96) = .68, p < .001$. Together, these patterns replicate previous work on children's number perception and demonstrate that children attended to and successfully understood the task. Additionally, they confirm that our manipulation of numeric ratio should also manipulate children's sense of certainty.

2.2.2. Post-choice certainty task

Because each trial consisted of a smaller and a larger ratio, we expected that children who attended to and compared two states of certainty would choose the larger (i.e., more certain) ratio more often than the smaller one. Consistent with this, 5, 6, and 7-year-olds showed this pattern and chose the more certain ratio more than 50% of the time (see Table 1 for means and tests against chance)[1] . We find these effects irrespective of the order in which children completed the tasks, making it unlikely that children relied on their memory of positive feedback from the Pre-Choice task to determine which question to answer. As a further examination of whether children chose trials based on their certainty, we examined whether children's choices actually reflected the trials that they answered correctly vs. incorrectly. Overall, children's accuracy was higher on the number discrimination trials that they kept during the certainty trials ($M = 84\%, SD = 14\%$), than on those they discarded ($M = 75\%, SD = 11\%$), $t(97) = 7.28, p < .001, d = 0.74$. This confirms that, for the majority of children in our sample, their choices in the task reflected a judicious strategy of choosing trials with the higher probability of success – i.e., trials with higher certainty.

[1] A small number of children ($n = 12$) adopted the opposite strategy, in which they consistently chose the smaller of the two ratios, often saying that they wished to challenge themselves. We report additional exploratory analyses on these children at the end of the Results section.

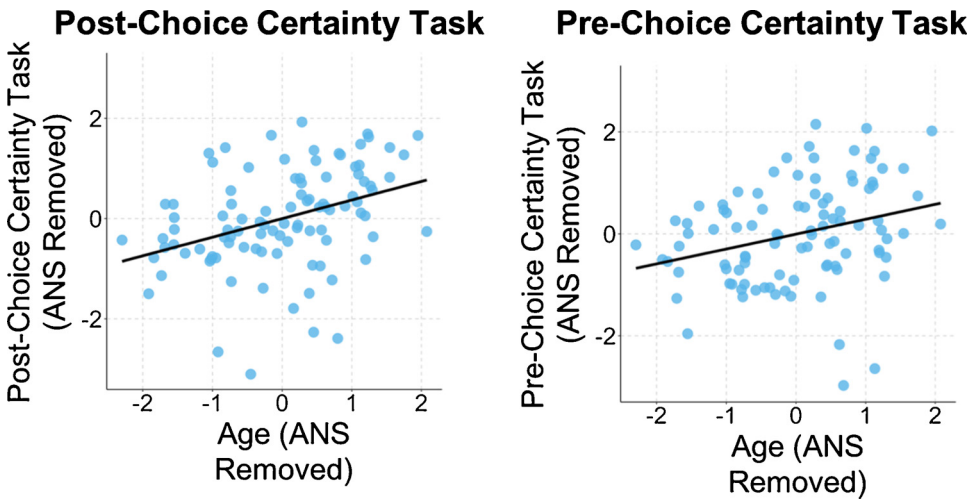

Fig. 3. Partial correlations between certainty accuracy and age, controlling for number discrimination (ANS) accuracy in Experiment 1.

Next, we turn to the central question of interest: which factors predict the development of children's ANS certainty? We found a strong correlation between children's choices on the Post-Choice Certainty Task and age, $r(96) = .40$; $p < .001$. Since the relative confidence task eliminates the need for criterion-setting, this result suggests that children's ANS certainty sensitivity develops independently of their criterion-setting abilities. We found this same result when examining the correlation between age and the ANS accuracy on chosen vs. unchosen trials, $F(4, 93) = 6.44$, $p < .001$; $\eta_p^2 = .22$.

But, could this age-related improvement simply be due to children's improving ANS precision? We found a trending correlation between children's ANS discrimination accuracy and their choice on the certainty task, $r(96) = .19$, $p = .066$, suggesting that the ANS contributes some variance to children's performance on the certainty task. However, adding ANS discrimination ability to a linear regression between certainty and age did not improve the model predicting certainty choice over age alone, $R^2_{Change} = .01$, $F(1, 95) = 1.37$, $p = .245$, $\beta_{Age} = .50$, $t(97) = 3.90$, $p < .001$, $\beta_{ANS} = -.15$, $t(97) = -1.17$, $p = .245$, $VIF = 1.84$ (see Fig. 3), suggesting that there are age-related improvements in certainty sensitivity independent of the underlying improvements in ANS representations themselves. But, do we observe identical results in the Pre-Choice Certainty task, in which children have to evaluate their certainty prospectively rather than retrospectively?

2.2.3. Pre-choice certainty task

As in the Post-Choice Certainty task, we found that children ages 5, 6, and 7 in the Pre-Choice Certainty Task chose the more certain ratio more than 50% of the time (see Table 1 for means and tests against chance)[2], and age correlated with the Pre-Choice accuracy, $r(97) = .47$, $p < .001$, suggesting that criterion-setting is not the only factor responsible for the development of ANS certainty. We also found a small age-related difference between the two Certainty tasks: as can be seen in Fig. 2, even 4-year-olds were able to select the more certain trials on the two largest (i.e., most disparate) metaratios, $M = 62.50$, $SD = 19.54$, $t(13) = 2.39$, $p = .033$, $d = 0.64$, despite not showing performance different from chance with all trials combined (see Table 1). Therefore, it is possible that young children's ANS certainty is so noisy and imprecise that they cannot reliably tell apart the metaratios we presented, but that they might succeed if given easier metaratios.

Replicating the Post-Choice results again, children's ANS discrimination performance and their choice of the more certain ratio also correlated, $r(95) = .40$, $p < .001$. However, adding number discrimination accuracy to a linear regression on choice of the more certain ratio did not explain any additional variability compared to age alone, $R^2_{Change} = .01$, $F(1, 94) = 1.47$, $p = .228$, $\beta_{Age} = .36$, $t(96) = 2.94$, $p = .004$, $\beta_{ANS} = .15$, $t(96) = 1.21$, $p = .228$, $VIF = 1.86$, see Fig. 3, suggesting that the development of sensitivity to certainty is not entirely driven by improvements in the underlying perceptual representations themselves, even in a prospective task with reduced cognitive demands.

2.2.4. Correlations between the tasks

Because the Pre- and Post-Choice Certainty versions differed in several ways, we performed two additional comparisons between tasks to confirm that both versions were measuring the same underlying ability. First, certainty accuracy on the two tasks (i.e., choosing the larger ratio) correlated even when controlling for age and number discrimination accuracy, $r(93) = .32$, $p = .002$. Second, children's accuracy on the ANS trials they expressed higher certainty in (trials they chose to answer in the Pre-Choice version, and trials they chose to keep in the Post-Choice version) were nearly identical (Pre-Choice: $M = 85\%$, $SD = 11\%$, Post-Choice: $M = 84\%$, $SD = 14\%$), $t(96) = 0.29$, $p = .771$. In fact, these two accuracies correlated even when controlling for age, $r(94) = .37$,

[2] As in the Post-Choice task, we found that a sample of children ($n = 11$) consistently chose the harder of the two trials. We report an exploratory analysis of these children at the end of the Results section.

$p < .001$, suggesting that children were trying to choose questions in both versions that maximized their chance of success.

Together, these results show that the Pre- and Post-Choice tasks both tapped into children's ANS certainty and that, furthermore, the development of children's ANS certainty sensitivity occurs independently of criterion-setting and individual and developmental differences in ANS acuity itself.

2.2.5. Metaratio effects

Because we presented children with 5 different metaratios – ratios *between* the two presented numerical ratios – we also examined whether children's choice of the more certain ratio changed as a function of the metaratio. Specifically, we predicted that certainty itself might be noisy and continuous and therefore subject to Weber's law (Weber, 1978), which would mean that children should be best at differentiating two states of certainty that are far apart (i.e. larger metaratios) than close together (Barthelmé & Mamassian, 2009).

Consistent with this, we find that children's accuracy and speed improved as the metaratio grew in both the Post-Choice task (Accuracy: $F(4, 388) = 25.55, p < .001, \eta_p^2 = .21$; Speed: $F(4, 388) = 4.17, p = .003, \eta_p^2 = .04$; see Fig. 2) and the Pre-Choice task (Accuracy: $F(4, 376) = 28.41, p < .001, \eta_p^2 = .23$; Speed: $F(2.67, 261.24) = 4.01, p = .011, \eta_p^2 = .04$; Fig. 2). Children's age (as a covariate) interacted with accuracy by metaratio in the Post-Choice task, $F(16, 372) = 3.01, p < .001, \eta_p^2 = .12$, consistent with the findings reported earlier that 3 and 4-year-olds did not choose the more certain ratio more than chance in this task. However, we do not find an interaction between metaratio and age as a covariate on children's choices in the Pre-Choice task, $F(16, 376) = 1.39, p = .145, \eta_p^2 = .06$, as even the youngest children in our sample chose the larger ratio above chance in this task given a large enough metaratio. These results remain qualitatively identical if we define metaratios in terms of the difference (rather than ratio) between the two ratios, and broadly suggest that children's representations of certainty are themselves subject to internal noise and are consistent with Weber's law (Weber, 1978).

2.2.6. Exploratory analysis of the "Opposite Strategy"

As noted above, we found that a small subset of children in both the Pre- and Post-Choice tasks consistently chose the trial they were *less* certain of ($n = 12$ in the Post-Choice Certainty task and $n = 11$ in the Pre-Choice Certainty task). These children are easily identified because their performance shows a *reversed* metaratio effect: the higher the difference between the two ratios, the more likely they were to choose the lower ratio trial (see also (see also Baer et al., 2018; Odic, Pietroski, Hunter, Lidz, & Halberda, 2013 for a mathematical model that tests which children show reverse ratio performance). Interestingly, we find that the probability of a child adopting such a strategy is not consistent across the two tasks, with only two children in the sample demonstrating this behavior in both the tasks.

In the analyses reported above, we left all of the children's data as-is. But, children who use this opposite strategy introduce statistical heterogeneity into the data, as their significantly below-chance performance leads to bimodality and higher variability, despite the fact that, in principle, their behavior is clearly indicating an ability to differentiate their two states of certainty. Thus, we performed two additional exploratory analyses: one with these children removed from the sample, and one with their performance mathematically transformed as a difference from 50%, in order to verify whether any of our results could be attributed to this subsample of children.

When removing these children from the sample, we still found a significant correlation between age and accuracy in the Post-Choice task, $r(85) = .52$; $p < .001$, and the Pre-Choice task, $r(87) = .53$; $p < .001$. Both of these remained significant even when controlling for individual differences in Number Discrimination accuracy, Post-Choice: $r(84) = .46$, $p < .001$; Pre-Choice: $r(84) = .43$, $p < .001$. We also analyzed our data when we mathematically transformed these children's data by taking the absolute difference in accuracy from 50% (this equates the performance of children who performed above and below 50% to the same degree, e.g., 75% and 25%, as they could both *discriminate* the two trials equally well, but reported their lower confidence choice). We once again found a significant correlation between age and accuracy on the Post-Choice task, $r(96) = .49$; $p < .001$, and Pre-Choice task, $r(97) = .50$; $p < .001$), even when controlling for Number Discrimination accuracy (Post-Choice: $r(94) = .41$, $p < .001$; Pre-Choice: $r(94) = .41$; $p < .001$). Together, both of these analyses support the conclusion that certainty sensitivity develops independently of criterion-setting and the ANS, even when the opposite strategy children are excluded or have their data transformed.

2.3. Discussion

In Experiment 1, we found that children's sensitivity to ANS certainty improves from age 3 to age 7, and that this is not fully explained by improving ANS precision or criterion-setting abilities. We also found that both of the versions of the certainty task – the Post-Choice Certainty task which asked children to evaluate their certainty *retrospectively* and the Pre-Choice Certainty task which asked them to evaluate their certainty *prospectively* – strongly correlated and both showed development independent of criterion-setting or ANS precision. Finally, and consistent with previous reports, we found evidence in both tasks that certainty decisions are metaratio-dependent: the larger the difference in certainty between the two trials, the more likely children were to identify the more certain trial.

At the same time, however, Experiment 1 has two limitations. First, in order to keep children motivated in the Pre-Choice task, we provided them with explicit feedback on their dot discrimination performance (though we gave them no feedback on the certainty portion of the task); but, in order to have children evaluate their certainty retrospectively, the Post-Choice task could not give children any feedback at all. One possibility, therefore, might be that children were trained to attend to their certainty signal throughout the course of the Pre-Choice task and could not attend to their certainty spontaneously without feedback.

The second limitation of Experiment 1 concerns our stimuli: while our ANS displays controlled for the cumulative surface area of the dots, they did not control for other non-numeric visual features that have sometimes been shown to influence children's performance. For example, Gebuis and Reynvoet (2012); see also Clayton, Gilmore, & Inglis, 2015; Szűcs et al., 2013) show that adult observers frequently select the side that has the higher convex hull (i.e., the largest contour drawn around the dots) rather than the side that is more numerous. One possible explanation for the continued development of certainty when controlling for the ANS, therefore, might be that children used distinct dimensions on the certainty and dot discrimination parts of the tasks (e.g., choosing certainty based on convex hull, but dot discrimination based on number, or cumulative area, etc.).

To rule out these two possibilities, in Experiment 2 we once again tested 3–7 year-old children on a Pre- and Post-Choice Certainty tasks with two major changes: neither task featured feedback, and the dot stimuli were created using the Gebuis and Reynvoet (2011) algorithm that controls for five different non-numeric features (cumulative area, convex hull, density, cumulative circumference, and cumulative diameter/additive area). If any of these factors is responsible for the positive results we found in Experiment 1, we should find that children's performance in Experiment 2 should be no different from chance.

3. Experiment 2

3.1. Method

3.1.1. Participants

Using the correlation between age and children's certainty judgments in Experiment 1 (r = .40, from the Post-Choice Task), we conducted a power analysis and determined that 61 participants would be required to replicate this effect with 90% power at α = .05. We recruited and tested 61 children aged 3–7 years (M = 5;6, range = 3;2–8;0, 32 girls) from the same area and in the same manner as Experiment 1. One child completed the Post-Choice version only, so his data was retained for analysis of the Post-Choice task and removed for all other analyses.

3.1.2. Materials and procedures

We used new number discrimination stimuli in this experiment that controlled for five non-numeric visual features: cumulative area, density, convex hull, cumulative diameter, and cumulative circumference. Stimuli were generated using a program designed by Gebuis and Reynvoet (2011), which overall balances the number of trials in which any of these dimensions correlate with the same answer as number. In other words, if children use any of these cues consistently, their number discrimination performance should be at chance. Note that for the very easiest ratios, the software cannot generate trials that have cumulative diameter and circumference in the opposite direction from number. To prevent these cues from being usable in children's certainty judgments, we matched each of the easiest ratio trials with a very difficult trial that had the cues correlated in the same direction (e.g., if cumulative circumference was a possible cue on a ratio 5.5, it was also a cue on the matched trial of 1.1), preventing children from using these cues to decide which trial they were more certain of. Each trial varied continuously in ratio from 1.04 to 5.5, binned into 7 groups: 1.05, 1.10, 1.35, 1.64, 2.05, 3.75, and 5.15. Using these new stimuli, we developed certainty pairs in the same way as Experiment 1, with metaratios of 1.25, 1.5, 3.0, 4.0, and 5.0. All other aspects of the materials were identical to Experiment 1.

The procedures were the same as in Experiment 1 with one change: children were not given feedback about their number discrimination performance by the computer in the Pre-Choice condition. Instead, to equate the use of feedback between the two versions, children were only given periodic neutral affirmations (e.g., "Okay!", "Alright", "Let's do another one!") equally in both the Pre-Choice and Post-Choice Certainty tasks, and only during the time between trials so that they could not interpret it as giving them any corrective feedback.

With these changes, children took 5.2 min on average to complete the Post-Choice task and 3.1 min on average to complete the Pre-Choice task.

3.2. Results

We found no effects of gender or order on children's performance, and so collapse across these variables for all analyses.

3.2.1. Number discrimination

Replicating Experiment 1 and previous work, children correctly answered 74% (SD = 8.16) of number discrimination questions, $t(60) = 23.36$, $p < .001$, $d = 2.99$. We also found ratio effect, with children performing more accurately, $F(4.23, 253.63) = 99.74$, $p < .001$, $\eta_p^2 = .62$, and faster, $F(4.64, 278.11) = 3.91$, $p = .003$, $\eta_p^2 = .06$, on the higher ratios (see Fig. 4). Children's accuracy also strongly correlated with age, $r(59) = .66$, $p < .001$. Thus, even when controlling for the five non-numeric visual features, children's performance was above chance and indicates that they relied on number.

3.2.2. Post-choice certainty task

Children aged 6 and 7 consistently chose the trials with larger ratios above chance rates (see Table 2 for means and t tests). As in Experiment 1, we found that children were more accurate on trials for which they indicated high certainty (M = 79.13, SD = 12.92), than trials they chose to discard (M = 69.67, SD = 8.77), $t(60) = 4.96$, $p < .001$, $d = 0.86$. And, as before, we found that a subset of children (n = 12) chose the opposite strategy of consistently choosing the lower certainty trial.

Children's certainty choice correlated with both age, $r(59) = .56$, $p < .001$, and ANS accuracy, $r(59) = .42$, $p = .001$,

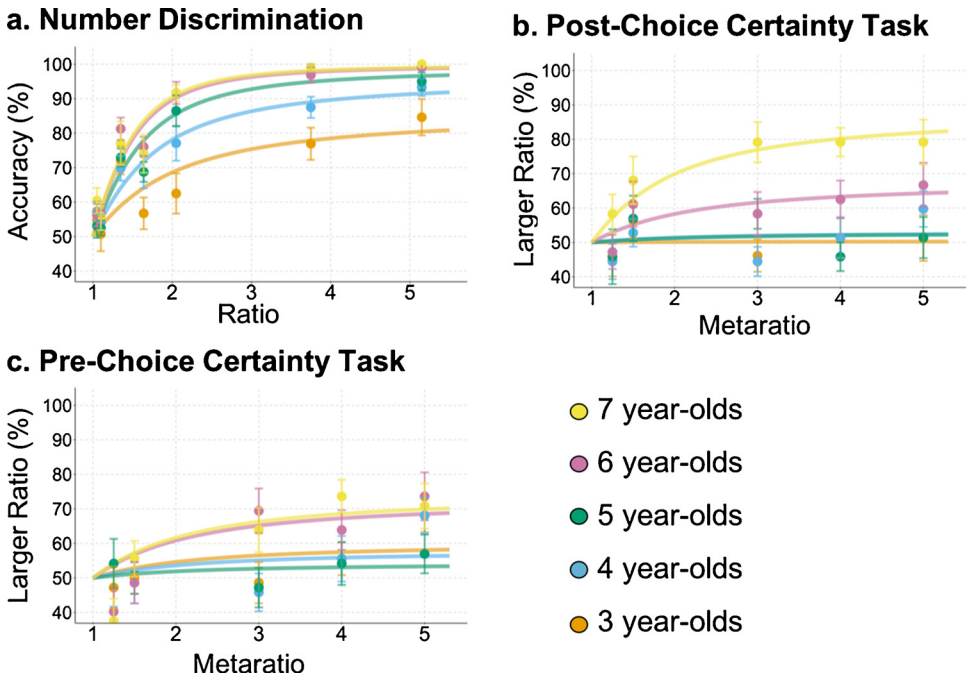

Fig. 4. Accuracy at each ratio on the number discrimination trials, and at each metaratio certainty trials in the Post-Choice and Pre-Choice Certainty Tasks in Experiment 2. Error bars represent 1 SE, and curves are estimated using a standard psychophysical model (see Odic, 2018).

Table 2
Sample sizes, means, tests against chance, and model fit estimates for the Number task, the Post-Choice version, and the Pre-Choice version in Experiment 2.

Age	*N*	% Correct (*SD*)	*t*	*p*	*d*
Number Discrimination Task					
Overall	61	74.34 (8.16)	23.36	< .001	2.99
3	13	65.64 (8.54)	6.60	< .001	1.83
4	12	72.64 (6.72)	11.67	< .001	3.37
5	12	75.28 (4.81)	18.20	< .001	5.26
6	12	79.58 (4.56)	22.49	< .001	6.49
7	12	79.58 (6.40)	16.01	< .001	4.62
Post-Choice Certainty Task					
Overall	61	56.72 (13.59)	3.86	< .001	0.49
3	13	50.00 (9.23)	0.00	1.00	0.00
4	12	50.56 (5.83)	0.33	.748	0.09
5	12	51.67 (9.59)	0.60	.559	0.17
6	12	59.17 (12.88)	2.47	.031	0.71
7	12	72.78 (12.55)	5.42	< .001	1.82
Pre-Choice Certainty Task					
Overall	60	55.33 (10.24)	4.04	< .001	0.52
3	12	53.89 (7.63)	1.77	.105	0.51
4	12	51.11 (9.36)	0.41	.689	0.12
5	12	52.22 (8.91)	0.86	.406	0.25
6	12	59.17 (10.26)	3.09	.010	0.89
7	12	60.28 (12.51)	2.85	.016	0.82

suggesting that their performance on the certainty portion was also not based on any of the five non-numeric visual features. And, once again, adding ANS discrimination ability to a linear regression between certainty and age did not improve the model predicting certainty choice over age alone, $R^2_{Change} = .00$, $F(1, 58) = 0.28$, $p = .599$, $\beta_{Age} = .51$, $t(57) = 3.57$, $p = .001$, $\beta_{ANS} = .08$, $t(57) = 0.53$, $p = .599$, $VIF = 1.77$ (see Fig. 5). The correlation between age and Post-Choice Certainty accuracy when controlling for Number Discrimination accuracy held if the 12 children using the opposite strategy were either removed $r(47) = .59$; $p < .001$, or had their performance mathematically transformed, $r(59) = .54$; $p < .001$.

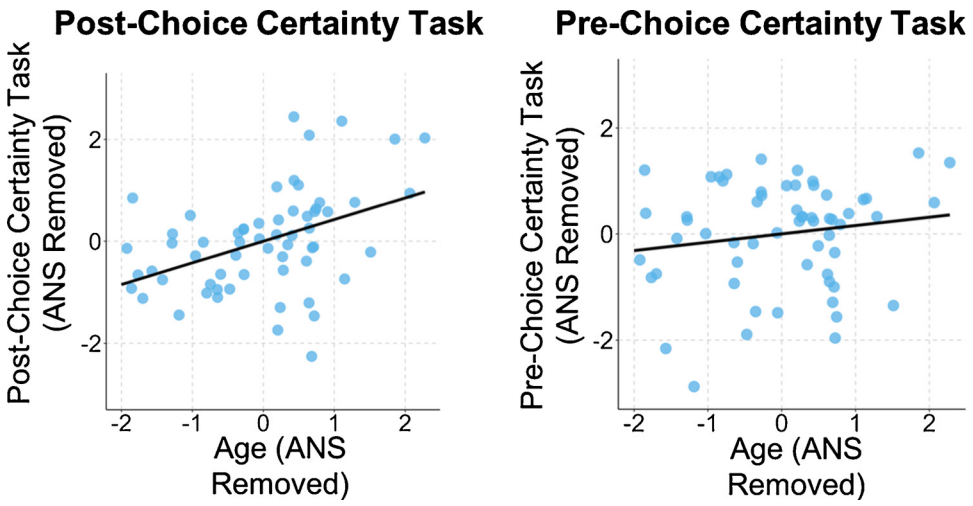

Fig. 5. Partial correlations between certainty accuracy and age, controlling for number discrimination (ANS) accuracy in Experiment 2.

3.2.3. Pre-choice certainty task

Replicating the Post-Choice results above, children aged 6 and 7 chose to answer trials with larger numerical ratios above chance rates (see Table 2 for means and t tests), indicating sensitivity to their certainty. And, as in the Post-Choice task, we found that a subsample of children (n = 10) consistently chose the lower certainty trial; only 4 children who went with this opposite strategy on both tasks.

Certainty choice on the Pre-Choice task correlated with age, $r(58) = .35$, $p = .007$, and ANS accuracy, $r(58) = .37$, $p = .003$. Adding ANS accuracy to the linear model predicting certainty choice did not improve the model over one with age alone, $R^2_{Change} = .04$, $F(1, 57) = 2.63$, $p = .111$, though we do note that it removed the effect of age when included, $\beta_{Age} = 0.18$, $t(56) = 1.14$, $p = .258$, $\beta_{ANS} = .26$, $t(56) = 1.62$, $p = .111$, $VIF = 1.71$ (see Fig. 5). Nevertheless, we found that age and certainty significantly correlated when controlling for ANS precision if the opposite strategy children were excluded, $r(48) = .44$, $p < .001$, or mathematically transformed as a difference from 50%, $r(58) = .42$, $p = .001$, consistent with the results of Experiment 1.

3.2.4. Correlations between the tasks

As in Experiment 1, children's performance on the Pre-Choice and Post-Choice versions correlated, $r(56) = .44$, $p = .001$, even when controlling for age and ANS precision. Moreover, children's accuracy on ANS trials they expressed higher certainty in were nearly identical (Pre-Choice: M = 81%, SD = 10%, Post-Choice: M = 79%, SD = 13%), $t(59) = 1.03$, $p = .307$, and were correlated even when controlling for age, $r(57) = .44$, $p = .001$.

3.2.5. Metaratio effects

As in Experiment 1, children were more likely to indicate high certainty in the larger of the two presented ratios when the metaratio between them was large, Pre-Choice: $F(3.31, 195.52) = 13.10$, $p < .001$, $\eta_p^2 = .18$, Post-Choice: $F(4, 60) = = 4.61$, $p = .001$, $\eta_p^2 = .07$. There were trending interactions between age (as a covariate) and metaratio predicting children's certainty choice in both versions, Pre-Choice: $F(3.34, 196.60) = 2.18$, $p = .083$, $\eta_p^2 = .04$, Post-Choice: $F(4, 236) = 2.11$, $p = .080$, $\eta_p^2 = .04$, where older children showed metaratio effects while younger children did not (see Fig. 4). We did not see any effect of metaratio for children's reaction times, Pre-Choice: $F(3.09, 182.03) = 1.32$, $p = .268$, $\eta_p^2 = .02$, Post-Choice: $F(3.33, 200.13) = 1.33$, $p = .263$, $\eta_p^2 = .02$.

4. General discussion

Young children's ANS representations provide them not only with an approximate sense of number, but also with a sense of certainty that improves with age: children become increasingly able to differentiate number discrimination trials that they believe they answered or could answer correctly vs. incorrectly. In two experiments, we tested whether this improving sensitivity in ANS certainty is accounted for by developmental improvements in calibration abilities, by the improving precision of children's ANS representations, or by improvements in children's more general ability to reason about their certainty. By testing 3–7-year-old children on two versions of the relative certainty task that directly measures sensitivity independent of criterion-setting, and by controlling for developmental improvements in children's ANS precision, we find that sensitivity in ANS certainty continues to develop until at least age 8. Importantly, these results hold even when feedback is entirely removed from the tasks, suggesting that children can access their certainty representations spontaneously, and when five non-numeric visual features, including density and convex hull, are controlled for. Our findings broadly replicate claims made in the literature that children improve at monitoring their certainty with age, and extend them by experimentally removing the influence of overconfidence bias and statistically removing the

influence of underlying ANS noise. They also contrast to previous reports that have argued that children's certainty develops primarily because of changes in criterion-setting (i.e., calibration; Salles et al., 2016).

What, then, are the additional factors contributing to the development of certainty sensitivity beyond calibration and ANS precision? Our results are consistent with the hypothesis that the improvement in children's certainty in ANS decisions is driven not by improvements in calibration or the ANS itself, but by improvements in the ability to reason about and represent perceptual certainty more generally. As one example, discussed in the Introduction, recent work has shown that certainty might act as a domain-general currency that bridges otherwise disparate and independent perceptual representations: children's certainty sensitivity in number, area, and emotion decisions is tightly linked, even though children's discrimination abilities in these three tasks are independent of each other (Baer et al., 2018). Similarly, adult observers are able to compare states of certainty across otherwise independent perceptual boundaries, such as visual vs. auditory trials or contrast vs. orientation. Moreover, we find here that children's ability to reason about their certainty is ratio-dependent, providing some empirical evidence that certainty is a continuous property that itself obeys Weber's law (Weber, 1978). Together, these findings are all consistent with the possibility that certainty is a type of domain-general magnitude itself, represented on a scale with noisy tuning curves akin to the representational format of the ANS (Halberda & Odic, 2014; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004). We see this as a fruitful avenue for future research.

Crucially, our claim is *not* that criterion-setting and underlying ANS precision do not contribute at all to children's development of certainty, but rather that these factors are not *sufficient* to explain certainty development by themselves. Some models suggest that certainty should be entirely a product of the low-level perceptual noise, and are not easily reconciled with our data (e.g., Maniscalco & Lau, 2012, 2014), while other models instead suggest that the certainty signal is aggregated from a variety of sources (e.g., Pleskac & Busemeyer, 2010 and Pouget et al., 2016). These sources are proposed to include the low-level perceptual noise (to some degree), but also the history of trials that the participant saw, their general belief about their ability, their estimate of how much attention they were paying on an individual trial, the strategy they are applying to the task, etc. (e.g., Koriat, 1993; Martí et al., 2018; Pleskac & Busemeyer, 2010; Pouget et al., 2016). Our data is most consistent with these aggregate models, though further work is required to understand precisely which sources of noise children draw upon when making certainty decisions in perceptual tasks.

A key open question is the extent to which perceptual confidence, as studied here, can be further generalized to even more global metacognitive abilities, such as children's ability to monitor their performance, understand appropriate strategies for specific tasks, know when to ask for help, etc. (e.g., Bellon, Fias, & De Smedt, 2019; Goupil, Romand-Monnier, & Kouider, 2016). For example, recent work has differentiated between broad metacognitive abilities like understanding how to apply strategies to tasks and math-specific numerical metacognitive abilities like assessing one's accuracy on addition problems, finding that numerical metacognition predicts math abilities in children, while broader metacognition does not (Bellon et al., 2019). Accordingly, an interesting extension of our work may be to examine how broadly the certainty we measure in a perceptual numerical task applies: whether to perceptual certainty tasks in general (e.g., emotion perception certainty), to number-relevant tasks in general (e.g., math performance), or perhaps even beyond (e.g., metamemory, or global strategies). We hope that the methodology established in the reported two experiments can be a launching platform for a deeper discussion about the relationship between perceptual confidence, mathematical metacognition, and metacognition more broadly.

Our study follows a tradition in the certainty monitoring literature of manipulating difficulty as a proxy for certainty because more difficult tasks intuitively should elicit less certainty. It may, therefore, be possible that children could be reasoning about the relative difficulty of the two trials (i.e., the *objective* probability of success, as indexed by the ratios of each trial; Nicholls, 1980; Nicholls & Miller, 1983), rather than their relative certainty in the tasks. However, we suspect that this is not the case for two reasons. First, consistent with the adult literature (e.g., Barthelmé & Mamassian, 2009; De Gardelle & Mamassian, 2014), children's choices tracked with their accuracy in the Post-Choice task: children were more likely to have correctly answered the trials they selected as higher in certainty than those they did not. Second, if children were making their decisions based solely on the ratios of the two presented trials without reasoning about their subjective certainty, we would expect that individual differences in ANS precision – which have previously been shown to correlate with and be instantiated in identical neural regions as ratio perception (Jacob & Nieder, 2009; Jacob, Vallentin, & Nieder, 2012; Matthews, Lewis, & Hubbard, 2016) – would account for any developmental improvements. Contrary to this, we found evidence for continuing development of certainty sensitivity when controlling for ANS precision in both the Pre- and Post-Choice tasks, suggesting that the ability to reason about ratios is not the only contributing factor to children's certainty task performance.

We set out to track age-related change in ANS certainty sensitivity in the preschool and early school years, but, in both Experiments, we did not find strong evidence that the youngest children in our sample were sensitive to certainty. This is in stark contrast to a growing body of work in toddlers and preschoolers which shows that children under age 5 *are* sensitive to certainty(e.g., Balcomb & Gerken, 2008; Call & Carpenter, 2001; Goupil & Kouider, 2016; Goupil et al., 2016; Lyons & Ghetti, 2011) and that certainty sensitivity develops and peaks by age 6 (Salles et al., 2016), prompting a question about why preschoolers in our sample did not show such sensitivity. One possibility is that the number discrimination task, which to our knowledge has only been used to elicit certainty in children aged 5 and older, does not elicit any sense of certainty in these younger children. However, we found that 4-year-olds showed above-chance performance on the two largest metaratios in Experiment 1, which might suggest that these younger children *are* capable of reasoning about certainty in this task but that the contrasts we used in our relative task were simply too close for young children to tell apart (much like infants can only discriminate large differences in number; Izard et al., 2009; Xu & Spelke, 2000). Therefore, consistent with the interpretation that certainty is represented on a continuous and noisy domain-general scale, perhaps young children can only discriminate large differences in certainty – larger than we presented in these tasks. At the same time, however, we failed to observe above-chance performance on the easiest ratios in Experiment 2 (thought this could be due to the lack of feedback), leaving the factors that lead to the youngest children's success on relative certainty tasks an open question. Future

work using this task with young children could make use of even larger metaratios or use different stimuli like area discrimination that children can discriminate more precisely (e.g., Odic, 2018) to test this interpretation.

Finally, it is important to note that although we have discussed our work in the context of developmental changes, our methodology was cross-sectional. While our experiments are primarily focused on how developmental differences in certainty sensitivity are not accounted for by ANS precision or criterion-setting, future work utilizing longitudinal designs would be better situated to understand the role of maturity vs. experience in accounting for the changes in sensitivity that we observed, as well as identifying whether or not development proceeds linearly.

In sum, children can reason about their certainty in a relative task, showing development in their precision with age. Age-related differences are not explained by children's numerical precision, suggesting an independent maturation process for certainty monitoring. We believe that this method of measuring individual differences opens up possibilities for deepening our understanding of certainty both in childhood and across many different populations.

Acknowledgements

This work was supported by an Insight Development Grant from the Social Sciences and Humanities Research Council of Canada (SSHRC) to DO, and a SSHRC Joseph-Armand Bombardier Canada Graduate Scholarship to CB. Thanks to members of the Centre for Cognitive Development (Andy Park, Kim Go, Mimi Zhang, Natasha Au, Stephanie Lee, Inderpreet Gill, Puja Malik) for their help with recruitment and data collection, and to the schools and families who participated. (note to editor: SSHRC awards do not have unique ID numbers).

References

Baer, C., Gill, I. K., & Odic, D. (2018). A domain-general sense of confidence in children. *Open Mind: Discoveries in Cognitive Science, 2*, 86–96. https://doi.org/10.1162/opmi_a_00020.

Balcomb, F. K., & Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science, 11*, 750–760. https://doi.org/10.1111/j.1467-7687.2008.00725.x.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology, 5*, e1000504. https://doi.org/10.1371/journal.pcbi.1000504.

Bellon, E., Fias, W., & De Smedt, B. (2019). More than number sense: The additional role of executive functions and metacognition in arithmetic. *Journal of Experimental Child Psychology, 182*, 38–60. https://doi.org/10.1016/j.jecp.2019.01.012.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436. https://doi.org/10.1163/156856897X00357.

Butterfield, E. C., Nelson, T. O., & Peck, V. (1988). Developmental aspects of the feeling of knowing. *Developmental Psychology, 24*, 654–663. https://doi.org/10.1037/0012-1649.24.5.654.

Call, J., & Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition, 3*, 207–220. https://doi.org/10.1007/s100710100078.

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

De Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PloS One, 11*, e0147901. https://doi.org/10.1371/journal.pone.0147901.

De Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science, 25*, 1286–1288. https://doi.org/10.1177/0956797614528956.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics, revised and updated edition.* USA: Oxford University Press.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*, 843–876. https://doi.org/10.3758/BF03196546.

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods, 43*, 981–986. https://doi.org/10.3758/s13428-011-0097-5.

Gebuis, T., & Reynvoet, B. (2012). Continuous visual properties explain neural responses to nonsymbolic number. *Psychophysiology, 49*, 1481–1491. https://doi.org/10.1111/j.1469-8986.2012.01461.x.

Goupil, L., & Kouider, S. (2016). Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Current Biology, 26*, 3038–3045. https://doi.org/10.1016/j.cub.2016.09.004.

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences, 113*, 3492–3496. https://doi.org/10.1073/pnas.1515129113.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Wiley.

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

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

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*, 665–668. https://doi.org/10.1038/nature07246.

Halberda, J., & Odic, D. (2014). The precision and internal confidence of our approximate number thoughts. *Evolutionary Origins and Early Development of Number Processing,* 305–333. https://doi.org/10.1016/B978-0-12-420133-0.00012-0.

Hembacher, E., & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*, 1768–1776. https://doi.org/10.1177/0956797614542273.

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

Jacob, S. N., & Nieder, A. (2009). Tuning to non-symbolic proportions in the human frontoparietal cortex. *The European Journal of Neuroscience, 30*, 1432–1442. https://doi.org/10.1111/j.1460-9568.2009.06932.x.

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: The brain's code for proportions. *Trends in Cognitive Sciences, 16*, 157–166. https://doi.org/10.1016/j.tics.2012.02.002.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639. https://doi.org/10.1037/0033-295X.100.4.609.

Lipowski, S. L., Merriman, W. E., & Dunlosky, J. (2013). Preschoolers can make highly accurate judgments of learning. *Developmental Psychology, 49*, 1505–1516. https://doi.org/10.1037/a0030614.

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development, 82*, 1778–1787. https://doi.org/10.1111/j.1467-

8624.2011.01649.x.
Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science, 2*, 459–481. https://doi.org/10.1146/annurev-vision-111815-114630.
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*, 422–430. https://doi.org/10.1016/j.concog.2011.09.021.
Maniscalco, B., & Lau, H. (2014). Signal Detection Theory analysis of type 1 and type 2 data: Meta-d′, response-specific Meta-d′, and the unequal variance SDT model. *The Cognitive Neuroscience of Metacognition,* 25–66. https://doi.org/10.1007/978-3-642-45190-4_3.
Martí, L., Mollica, F., Piantadosi, S., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind: Discoveries in Cognitive Science, 2*, 47–60. https://doi.org/10.1162/opmi_a_00017.
Matthews, P. G., Lewis, M. R., & Hubbard, E. M. (2016). Individual differences in non-symbolic ratio processing predict symbolic math performance. *Psychological Science, 27*, 191–202. https://doi.org/10.1177/0956797615617799.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133. https://doi.org/10.1037/0033-2909.95.1.109.
Nicholls, J. G. (1980). The development of the concept of difficulty. *Merrill-Palmer Quarterly, 26*, 271–281.
Nicholls, J. G., & Miller, A. T. (1983). The differentiation of the concepts of difficulty and ability. *Child Development, 54*, 951–959. https://doi.org/10.2307/1129899.
Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science, 21*, e12533. https://doi.org/10.1111/desc.12533.
Odic, D., Pietroski, P., Hunter, T., Lidz, J., & Halberda, J. (2013). Young children's understanding of "more" and discrimination of number and surface area. *Journal of Experimental Psychology Learning, Memory, and Cognition, 39*, 451–461. https://doi.org/10.1037/a0028874.
Odic, D., & Starr, A. (2018). An introduction to the Approximate Number System. *Child Development Perspectives, 12*, 223–229. https://doi.org/10.1111/cdep.12288.
O'Leary, A. P., & Sloutsky, V. M. (2017). Carving metacognition at its joints: Protracted development of component processes. *Child Development, 88*, 1015–1032. https://doi.org/10.1111/cdev.12644.
Piazza, M., De Feo, V., Panzeri, S., & Dehaene, S. (2018). Learning to focus on number. *Cognition, 181*, 35–45. https://doi.org/10.1016/j.cognition.2018.07.011.
Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44*, 547–555. https://doi.org/10.1016/j.neuron.2004.10.014.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864–901. https://doi.org/10.1037/a0019737.
Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience, 19*, 366–374. https://doi.org/10.1038/nn.4240.
Salles, A., Ais, J., Semelman, M., Sigman, M., & Calero, C. I. (2016). The metacognitive abilities of children and adults. *Cognitive Development, 40*, 101–110. https://doi.org/10.1016/j.cogdev.2016.08.009.
Smith, J. D., Beran, M. J., Couchman, J. J., & Coutinho, M. V. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review, 15*, 679–691. https://doi.org/10.3758/PBR.15.4.679.
Szűcs, D., Nobes, A., Devine, A., Gabriel, F. C., & Gebuis, T. (2013). Visual stimulus parameters seriously compromise the measurement of approximate number system acuity and comparative effects between adults and children. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00444.
Tamis-LeMonda, C. S., Adolph, K. E., Lobo, S. A., Karasik, L. B., Ishak, S., & Dimitropoulou, K. A. (2008). When infants take mothers' advice: 18-month-olds integrate perceptual and social information to guide motor action. *Developmental Psychology, 44*, 734–746. https://doi.org/10.1037/0012-1649.44.3.734.
Vallortigara, G. (2017). An animal's sense of number. In J. Adams, P. Barmby, & A. Mesouid (Eds.). *The nature and development of mathematics: Cross disciplinary perspectives on cognition, learning and culture* (pp. 43–66). London, UK: Taylor & Francis.
van Loon, M., de Bruin, A., Leppink, J., & Roebers, C. (2017). Why are children overconfident? Developmental differences in the implementation of accessibility cues when judging concept learning. *Journal of Experimental Child Psychology, 158*, 77–94. https://doi.org/10.1016/j.jecp.2017.01.008.
Vo, V. A., Li, R., Kornell, N., Pouget, A., & Cantlon, J. F. (2014). Young children bet on their numerical skills metacognition in the numerical domain. *Psychological Science, 25*, 1712–1721. https://doi.org/10.1177/0956797614538458.
Weber, E. H. (1978). *The sense of touch* (1st ed.). London, UK: Academic Press for Experimental Psychology Society.
Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74*, B1–B11. https://doi.org/10.1016/S0010-0277(99)00066-9.