DEVELOPING A SENSE OF CERTAINTY

by

Carolyn Elizabeth Baer

B.A., The University of Waterloo, 2014
M.A.Sc., The University of Waterloo, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
in
THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Developing a Sense of Certainty

submitted by Carolyn Elizabeth Baer in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Psychology

Examining Committee:

Dr. Darko Odic, UBC Department of Psychology

Supervisor

Dr. Susan Birch, UBC Department of Psychology

Supervisory Committee Member

Dr. Rebecca Todd, UBC Department of Psychology

University Examiner

Dr. Anne-Michelle Tessier, UBC Department of Linguistics

University Examiner

Additional Supervisory Committee Members:

Dr. Toni Schmader, UBC Department of Psychology

Supervisory Committee Member

Abstract

In a noisy world filled with confusion, humans need a toolkit of skills to help discern fact from fiction. In this dissertation, I explore one such tool and its development in childhood: metacognitive reasoning about confidence. In 7 studies, I investigate how children reason about the strength of subjective information, uncovering the properties of childhood metacognition and using its development as a tool to learn about metacognition more broadly.

Fueling this research are two families of theoretical accounts: one which conceptualizes confidence as a direct readout of decision noise (Direct accounts), and one where confidence is a combination of information from multiple sources (Inferential accounts). These two accounts make divergent predictions about two properties of metacognition: (1) how tightly bound confidence is to the underlying decisions it evaluates, and (2) how broadly is confidence represented in the mind. I investigate these questions using a developmental lens for testing between these theories.

These studies present a force-choice method of measuring children's sensitivity to confidence by asking how closely they can tell apart two states of confidence. This method of assessing confidence allows me to narrow in on the properties of metacognition that develop independently of children's overconfidence biases and developing linguistic knowledge. In Chapter 2, I use this measure to look for developmental change associated with confidence judgments when controlling for decision noise, finding age-related change consistent with the Inferential accounts. In Chapter 3, I test whether children reason about confidence using encapsulated systems or a broader metacognitive system, and probe whether these judgments share a unit of representation, finding evidence for both within the domain of perceptual judgments as predicted again by Inferential accounts. In Chapter 4, I investigate whether confidence is processed so broadly as to include reasoning about others' abilities, but do not find strong evidence of this, suggesting a limit on the generality of confidence processing.

All together, this dissertation shows that far from being subject to the whims of others, children possess a sense of confidence that combines multiple sources in information to create broadly-usable assessments of truth in the world.

Lay Summary

The world can be a confusing place, but we possess a sense of confidence to help determine which information is reliable. In this dissertation, I explore how this sense of confidence develops – uncovering its core cognitive properties and its use in early childhood learning. In 7 studies, I test competing theories about the fundamental nature of confidence reasoning in humans by investigating what information children use to generate feelings of confidence, and how broadly confidence is represented in the mind. Overall, I find that children possess a sense of confidence that combines multiple sources in information to create broadly-usable assessments of truth in the world. These studies are informative both to the study of confidence processing and to the broader understanding of children's learning.

Preface

I am the primary author of the work presented in this dissertation, which was all done under the supervision of Dr. Darko Odic. In all cases, I designed the studies, collected the data or supervised data collection, analyzed the results, and drafted the manuscripts. D.O. provided critical feedback on study design, analysis, interpretation of the results, and the written manuscripts. All studies in this dissertation were approved by the UBC Behavioural Research Ethics Board under certificate H14-01984.

Chapter 2:

A version of this work has been published: Baer, C. & Odic, D. (2019). Certainty in numerical judgments develops independently of the Approximate Number System. *Cognitive Development, 52*, 100817. doi:10.1016/j.cogdev.2019.100817

Chapter 3:

A version of Study 3 has been published: Baer, C., Gill, I.K., & Odic, D. (2018). A domain-general sense of confidence in children. *Open Mind: Discoveries in Cognitive Science, 2*, 208-218. doi:10.1162/opmi_a_00020

I.K.G. contributed to study design and data collection.

A version of Study 4 has been published: Baer, C. & Odic, D. (2020). Children flexibly compare their perceptual certainty within and across perceptual domains. *Developmental Psychology, 56*, 2095-2101. doi:10.1037/dev0001100

Chapter 4:

A version of this work is being revised for publication: Baer, C., Malik, P., & Odic, D. (in revision). Are children's judgments of another's accuracy linked to their metacognitive confidence judgments?

P.M. contributed to study design and data collection.

Table of Contents

List of Tables

List of Figures

Acknowledgements

With this dissertation, I finally complete grade 22 (and a half, thanks to COVID-19). That is 22 years (plus a few before that) where I have relied on so many people to guide my learning, to keep me motivated, to keep me fed, and so much more. It's a bit overwhelming to think about – and that comes from someone who has spent the last 5 years thinking about uncertainty. So here goes:

To the 593 children who became scientists for a day to make these studies happen, the 3863 children who helped us develop these studies and participate in the many others I've run, and their caregivers who saw the value in research, thank you. I hope that through 'playing computer games' you have seen how cool science can really be.

To the 80 schools and daycares, with hundreds of teachers, staff, and directors who facilitated this work alongside their already busy workloads, thank you. Discussing these projects with you has been so helpful in bringing this work back to the big reason I did it in the first place.

To the Social Sciences and Humanities Research Council of Canada and the endowment of Elizabeth Young Lacey through UBC who provided financial support for the last 5 years, thank you.

To the Musqueam peoples for welcoming us to your traditional, ancestral, and unceded lands, thank you.

To my many academic mentors over the years, notably Sue and Ori, who fostered the budding scientist in me and have continued to support me throughout my graduate career, thank you. I am so lucky to be part of an intellectual community with people who are simultaneously as intelligent, generous, and caring as you.

To the 23 research assistants and interns I have supervised at the Centre for Cognitive Development, plus the dozens more who worked in the lab with me, who donated their time to making this research happen, thank you. Your enthusiasm, curiosity, skepticism, optimism, and persistence has been so inspiring, and working with all of you has been one of the most rewarding experiences of my life.

To my lab family, Denny and Cory, who spent hours with me in and out of the lab celebrating the wins and commiserating the losses, thank you. You have both been such an inspiration and such amazing partners in crime.

To my family, Ron, Cathy, Angela, my grandparents, and extended relatives, who instilled a love of learning, always made education a priority, and made sure I was well taken care of no matter what, thank you. It's hard to express just how lucky I feel to have been born into such a wonderful support system.

To Alex, who moved across a (very big) country, made sure I was fed and watered, and pushed me to take risks and stay social, thank you. Your partnership has been the perfect antidote to the crazy ups and downs of grad school.

To Darko, who has taught me so much about science, communication, leadership, and citizenship, and who somehow seemed to know exactly what to say in every situation both to challenge me as an academic and to support me as a human, thank you. I feel so honoured to have you as a mentor.

And finally, to all the supporters over the years who have given me words of encouragement and kept me sane, thank you. From my VS crew, to my fellow graduate students and post docs at UBC, to colleagues around the world, to my Waseosa family, to the KMSB, to friends old and new, you have all enriched my life so much, and I'm so grateful for each and every one of you.

Thank you.

Chapter 1: Introduction

There is no doubt that the world can be a confusing place. We're faced with constant decisions about which route to take home or what to say to an upset friend. Our minds play tricks on us when ordinary eye blinks seem to be flickering lights, or a terrifying creature in a dark alleyway turns out to be garbage in the wind. More than that, we see the sun disappear every night and hope that it returns the next day, and we go back to a favourite restaurant because we expect the food to be just as good the second time, but neither of these is truly guaranteed. How can we form accurate predictions and make sound decisions amid all this confusion?

As adults, we have a sophisticated cognitive toolbox that helps us navigate the uncertain landscape we face. Over many years, we amass knowledge about the physical laws that govern our planet and the social norms that govern our culture. We experience the sun rising every day and learn that sometimes our senses are not to be trusted. Using this wide array of knowledge and experience about the causes, consequences, and predictability of events, we can tame the uncertainty we face and make sense of the world to make faster and more accurate decisions.

However, this knowledge and experience takes time to acquire and is itself dependent on trustworthy information about the world and its many moving parts. What, then, do infants and children do to understand the world around them? Children don't have the luxury of years of experience to form accurate expectations about what causes noises in alleyways (after all, they don't even have the luxury of fully developed perceptual and cognitive systems). And yet, children quite universally grow up to be adults who possess this sophisticated cognitive toolbox.

Many prominent theories have uncovered various tools that children use to make sense of the world. As a few of many examples, infants have a very early developing set of intuitions about the world, including the core properties of objects, numbers, and agents (Carey, 2009; Spelke & Kinzler, 2007). Children can also rapidly generalize new information from individuals to kinds, inferring deep, essential properties of categories from as little as one exposure (Csibra & Gergely, 2009; Gelman, 2003). And, as part of a social species, children are uniquely attuned to the information provided by others, trusting others to be truthful when they communicate (Harris, 2012).

Armed with all these abilities, we might get the impression that children are injudicious learners, absorbing whatever knowledge they can as quickly as possible, thereby subjecting

themselves to the whims of others who may intentionally or accidentally mislead them (Jaswal, 2010). But in contrast, we see ample evidence that children are *rational* learners. For instance, infants and children throughout development readily track the statistics of their environment, making sophisticated predictions about the patterns of future events based on past events (Saffran et al., 1996). By the early preschool years, children are skeptical of claims made by previously unreliable persons (Kidd et al., 2013; Mills, 2013), and preferentially seek information from reliable teachers (Poulin-Dubois & Brosseau-Liard, 2016). In fact, children appear to rationally weigh the strength of their prior knowledge and experience against new evidence, sticking with prior knowledge when new evidence seems unreliable (Gopnik & Bonawitz, 2015; Sobel & Kushnir, 2013).

Central to these claims of rationality in children's learning is an ability to evaluate the strength of evidence – both for incoming evidence and for one's own prior knowledge. In some cases, like reasoning about the potential causes of an event, the strength of evidence is reduced to an objective probability or likelihood: if event A always occurs after event B, but 50% of the time after event C, then we can calculate that there is much stronger evidence for B as the cause of A (Denison & Xu, 2019; Gopnik et al., 2004). A considerable case has been built for this kind of objective probabilistic reasoning in infants and children using evidence from causal reasoning (Gopnik & Bonawitz, 2015; Gopnik & Schulz, 2004), demonstrating that children possess the capability to both evaluate probability and compare it sensibly.

In other cases, children must reason about the strength of *subjective* evidence, say the likelihood they know the meaning of the word "science" or whether they will be capable of answering this new homework question. These cases have been the target of research into *metacognitive* processes: cognition that is about our own cognition (Flavell, 1979; Mamassian, 2016; Pouget et al., 2016). More specifically, we experience the strength of subjective evidence through feelings of *confidence*, or a sense of *certainty* – subjective assessments of how likely we are to be correct about our thoughts, decisions, and knowledge. Confidence then functions in much the same way probabilities do for objective evidence, providing graded assessments of evidence strength (Pouget et al., 2016).

But despite the apparent utility of confidence in establishing rational thought, confidence itself has remained somewhat of an unknown quantity in cognitive science. By virtue of being subjective, confidence is not a property of events in the world, it is a property of the mind – and a

property of that mind's representation *of itself* at that. Confidence also does not occur in isolation; we are always certain *about* a decision or a piece of knowledge. How then do our minds generate these feelings of confidence? Do we have dedicated cognitive resources for reasoning about confidence separately from the cognition it evaluates, and if so, how are these resources involved in rationally evaluating the strength of incoming evidence? And, in particular, how might developmental change impact these representations of confidence? In this dissertation, I examine these questions by exploring children's metacognitive reasoning about confidence, learning why metacognitive reasoning develops and how broadly it is computed.

Do Children Have a Sense of Confidence?

Anyone who has interacted with children (particularly extraverted, Western children) knows first-hand that children are incredibly unreliable sources when it comes to reporting on their own capabilities. From the wishful thinker who claims they can fly, to the defiant child who insists they "didn't hear" the request to begin clean-up, it's not hard to find children who have expressed high confidence for something they are ultimately and often transparently inaccurate about. For years, this belief was the dominant view of metacognitive ability in childhood: only by middle childhood (around age 8) were children found capable of accurately reporting on their confidence (Flavell, 1979).

However, this view has changed dramatically in the past few decades, as researchers have adapted adult-oriented measures for use with children. Rating scales from 0-10 or estimates of percentages are unreasonable measures to use with young children given the late emergence of understanding numbers precisely (Le Corre & Carey, 2007). One alternative measure developed for use with early school-aged children asks for a response on a "thermometer," a sliding scale from high to low confidence (van Loon et al., 2013; van Loon, Destan, et al., 2017). Or, one now popularly-used method of obtaining confidence reports from children uses a 3-point scale accompanied by pictures of a child expressing high, medium, and low confidence (Hembacher & Ghetti, 2014). To help children understand that the pictures or endpoints of the thermometer refer to states of confidence, the experimenter provides the children with corrective feedback on the proper use of the scale, including drawing the child's attention to their own behavioural indices of confidence like response speed and facial expression (Lyons & Ghetti, 2011). Using these measures, children as young as 3 years old provided higher confidence ratings on accurate than

inaccurate answers (Hembacher & Ghetti, 2014; Lyons & Ghetti, 2011), showing the key signature of metacognitive awareness before children enter formal schooling.

In verbal-report measures of metacognition where children must either explicitly communicate their knowledge state to an experimenter (e.g., Rohwer et al., 2012) or indicate their confidence on a simple scale (e.g., Hembacher & Ghetti, 2014; Lyons & Ghetti, 2011), the child's language understanding could explain development in confidence monitoring. That is, if a child does not have the language to communicate their uncertainty (e.g., does not understand the words *sure* or *know*), it will appear as though they cannot reflect on uncertainty.[1] But, there are numerous responses to uncertainty that do not rely on language but nonetheless reflect an evaluation of its strength: we will wait in a long line only if we are sure of eventually getting to the end of it, and we will seek help from others when we feel that we cannot find an answer ourselves.

Taking methods from the study of animal metacognition that rely on these behavioural predictions as inspiration (for review, see Kepecs & Mainen, 2012), developmental researchers have now found evidence of sensible metacognitive monitoring even by infants (Goupil et al., 2016; Goupil & Kouider, 2016; Kim & Kwak, 2011). For instance, when 20-month-old infants needed to locate a hidden object, they strategically sought help from their caregiver when they were unlikely to remember its location but not when they could easily remember (Goupil et al., 2016). Similarly, 3-year-old preschoolers opted out of answering questions that were difficult for them but not easier questions (Balcomb & Gerken, 2008; Bernard et al., 2015). Neural and physiological evidence also corroborates these findings: three-year-olds' pupil sizes increased in response to uncertainty (Paulus et al., 2013), and infants' brains as young as 12 months showed error-related negativity (an electrophysiological signature for detecting inaccuracies) following incorrect choices (Goupil & Kouider, 2016).

[1] One assumption held by some metacognition researchers and philosophers throughout the years is that confidence is an explicitly reportable output of metacognitive processes, computed with full conscious awareness (Carruthers, 2008; Perner, 2012). Other theorists argue for a separate classification of the metacognitive processes underlying these behaviours, referring to some as "implicit" metacognition (because present methods cannot confirm whether these decisions rely on "explicit" conscious reasoning) or "procedural" metacognition (referring to its use in guiding behaviour; Proust, 2007). However, these definitions presuppose that there are meaningful differences between the metacognitive processing in each task rather than testing for such differences empirically – it may actually be the case that *all* 'metacognitive' processing is computed 'implicitly' and conscious access to it is simply epiphenomenal. Regardless, the work here is interested in uncovering these processes rather than defining them, so I avoid the use of these terms.

In fact, there is now also ample evidence for diversity in children's metacognitive reasoning with minimal linguistic demands. In addition to seeking help from others or searching for more information (Call & Carpenter, 2001; Coughlin et al., 2015), children by age 5 use evaluations of their confidence to maximize a reward payout, placing high bets on items they are likely to answer correctly and low bets on incorrect items (Hembacher & Ghetti, 2013; Salles et al., 2016; Vo et al., 2014). Similarly, children at age 3-4 strategically exclude answers from evaluation when they are unsure of their accuracy (Hembacher & Ghetti, 2014; Lyons & Ghetti, 2013), and by age 9 will devote more study time to difficult unrelated word pairs (Lockl & Schneider, 2004). These metacognitive measures have been used for many kinds of decisions, from evaluations of memory (e.g., Balcomb & Gerken, 2008; Hembacher & Ghetti, 2013; Lockl & Schneider, 2004) to interpreting perceptual features (e.g., Beran et al., 2012; Lyons & Ghetti, 2011; Vo et al., 2014) to evaluating the optimal use of cognitive strategies (e.g., Geurten et al., 2018). And, importantly, many studies have reported correlations between these behavioural methods of measuring metacognition and children's explicit confidence reports (Coughlin et al., 2015; Hembacher & Ghetti, 2014; Lyons & Ghetti, 2013; Roebers et al., 2019), suggesting that the different measures tap into the same metacognitive ability. While this body of knowledge provides consistent evidence that children and infants *have* metacognitive abilities, it so far has done little to answer our central question: what *is* subjective confidence? In other words, what information do children recruit when they are reasoning about their confidence?

What is Subjective Confidence?

From our own introspective experiences, we know that confidence acts like a litmus test as we go about our daily activities, signalling when we should trust our experiences and when we're missing something and need to be skeptical. These experiences are largely well-founded. Across a vast collection of studies, confidence ratings on arbitrary scales (whether that be 0-10, 0%-100%, or 'low, medium, and high') overwhelmingly tend to correspond to the subject's ultimate accuracy (Pleskac & Busemeyer, 2010; Rahnev et al., 2020), signalling that human adults are introspectively attending to the reliability of their own cognition. For this reason, many domains of psychological research frequently use confidence judgments to help understand the human experience, even in the early days of psychophysical research (e.g., Pierce & Jastrow, 1884).

Pinning down exactly what confidence is and how it is computed, though, has proven difficult. As one example, the term ‘metacognition’ is frequently used to describe the process of computing confidence and reasoning about certainty (e.g., Fleming & Lau, 2014; Rouault et al., 2018; and this dissertation), but this term is also used by researchers interested in a whole host of self-regulatory behaviours and cognitive monitoring processes (Flavell, 1979). The knowledge that studying will improve your test performance, for instance, is considered metacognitive even though such reasoning does not involve actively monitoring one’s own thoughts (Veenman et al., 2006). These different processes, while all 'metacognitive' in the sense that they are cognition that takes as its object other cognition, likely reflect different cognitive processes – one that is consciously introspective and fully aware of itself (e.g., making study plans), and another that does not have to be (e.g., experiencing and reasoning about uncertainty). Researchers interested in one component of metacognitive reasoning, like confidence representations, must therefore filter through findings that likely reflect entirely different cognitive processes.

Even once we narrow down our definition of metacognition to the evaluation of confidence and certainty, we see signatures of the multi-faceted approach to metacognitive research in the proposals for cognitive processes underlying the computation and use of confidence representations. From the vast theories spanning philosophy, perception, memory, decision-making, neuroscience, and development, two broad types of theories dominate: Direct and Inferential. Importantly, each of these broad families makes distinct predictions about how confidence relates to other cognitive processes throughout development.

Direct Accounts

We have long known that the brain is directly responsible for our thoughts, which has led to a challenge of determining exactly how we can account for the complex reasoning humans do with a series of biological functions. The vantage point taken by the Direct family of accounts is that ultimately observers estimate their confidence based on the *local* properties of their task-specific representations (e.g., the amount of perceptual noise an observer is experiencing when representing the world). For example, if a participant is trying to identify a letter printed on a page, both their ability to identify the letter and to report their confidence ultimately stems from the sharpness of their vision: when there is more imprecision in the visual representation,

confidence should directly and proportionally decrease as well. Below, I discuss this family of accounts through two related theories: Signal-Detection and the Accumulator model.

The Signal-Detection Theory Account. For many years, confidence has largely been studied within the context of Signal Detection Theory (SDT; Galvin et al., 2003; Maniscalco & Lau, 2012; Pouget et al., 2016). One of the fundamental tenets of SDT is that our representations are *noisy*: we experience a degree of uncertainty around every percept (e.g., what we perceive to be 300 words on a page could also be 237 or 413; Dehaene, 2011; Green & Swets, 1988). In this sense, one conceptualization of confidence is that it is a direct readout of this noise (Pouget et al., 2016). For example, let's imagine that we estimate the number of words on this page 100 times and draw a histogram based on our estimates. What we would end up with is a roughly Gaussian ('normal') curve: 300 would be the most frequent answer and estimates like 237 and 413 would appear relatively infrequently towards the tails. Confidence in this example reflects the standard deviation of the curve (and indeed, this is what we use in statistics to calculate "Confidence Intervals").

To illustrate how we can then use this standard deviation to extract a single estimate of confidence, it is helpful to label decisions about the world as *Type 1* decisions (e.g., which is brighter, which is more numerous, etc.), and decisions about confidence as *Type 2* decisions (e.g., ratings scales, decisions to opt-out, etc.; Clarke et al., 1959; Galvin et al., 2003). When we make a Type 1 decision, we ask whether our noisy representation is above a threshold, or is generally higher than another noisy representation (the way a one-sample t test compares a mean against a single value and an independent samples t test compares two means against each other). When we make a Type 2 decision, on the other hand, we examine the evidence in favour of our chosen option and determine whether it also exceeds a threshold of confidence (the way those same t tests output p values or effect sizes based on the standard deviations to tell us how meaningful or large a difference in means is).

Thus, a critical property of SDT models is that they treat confidence as deterministic. The confidence report is a direct reflection of (1) the levels of underlying noise (e.g., the standard deviation), (2) the placement of the decision threshold (e.g., the null hypothesis), and (3) the placement of the confidence threshold (e.g., the critical p value or effect size). Thus, a clear prediction of this account is that Type 2 representations should be directly estimated from Type 1 representations (which are the direct result of (1) and (2)) if we can properly account for

differences in threshold placement (Maniscalco & Lau, 2012). Taking this one step further, confidence reasoning critically relies on the noise of Type 1 representations, but the noise of Type 1 representations is specific to the task being performed. For instance, when our senses interpret the world, they do so in different units: when listening to somebody speak, vision might provide information about the movement of dark and light patterns of the lips, while audition interprets the pressure of sound waves in the room. This leads to a second prediction of this account: Type 2 decisions should be unrelated across tasks that generate different Type 1 noise (De Gardelle & Mamassian, 2014).

The perspective of the SDT account is mirrored by two related accounts. In the study of *metamemory* (the way in which we reason about the reliability of our memories; Dunlosky & Bjork, 2008; Nelson & Narens, 1990), a nearly identical model was proposed arguing that we monitor our own memory states through *direct access*: our metamemories are entirely determined by the same information used to make memory judgments (Koriat, 1993; Van Zandt, 2000). And, in the study of multisensory integration, computational models have shown that observers near-optimally integrate visual and auditory information by weighing each signal's perceptual precision – the inverse of SDT noise (Alais & Burr, 2004; Ernst & Banks, 2002). When judging spatial location, for instance, when there is an imprecise auditory signal, participants rely more heavily on using the visual signal to determine the location of an object (Alais & Burr, 2004). Crucially, even in the multisensory integration accounts, the only way in which confidence is computed across domains is by comparing information over identical Type 1 units – like spatial position generated independently by our visual and auditory processing. Confidence can therefore only be compared and integrated for domains where multiple systems code for the same information. Thus, just like classic SDT theories, a subject could directly read-out the weight of the integrated signal, predicting strong associations between Type 1 and Type 2 reasoning, but not between multiple Type 2 decisions that stretch across distinct domains.

Because confidence calculations are performed deterministically, an interesting feature of this family of theories is that they should operate identically in children as they do in adults (Salles et al., 2016). The only potential sources for developmental change, then, would be the noise of the representations themselves or the placement of the Type 1 and Type 2 thresholds (decision and confidence). In other words, if each of the three components of the confidence

computation can be fully accounted for (e.g., Type 1 bias, Type 2 bias, and Type 1/2 noise), there should be no remaining development in children's metacognitive ability.

The Accumulator Account. Classic SDT models, such as those discussed above, do not model the effect of time, instead assuming that our decisions are made instantly and only once. Accumulator models, on the other hand, directly incorporate time into SDT models, viewing decision-making as an *evidence accumulation* process (De Martino et al., 2013; Kiani & Shadlen, 2009; Pleskac & Busemeyer, 2010). In turn, accumulator models critically point out that time often strongly correlates with decision accuracy and with decision confidence (Rahnev et al., 2020), something that classic SDT models do not consider. In these models, decisions unfold over time by gaining new evidence at regular intervals for each option in the decision. As an example of one such model, known as a *random walk,* picture a tug-of-war occurring between team A and team B where one team must pull the centre of the rope past a threshold on their side. Both teams are capable of pulling the rope toward their side, but the winning team is ultimately the one who most *consistently* pulls the rope. Analogously, evidence for options A and B accumulates over time until one surpasses a decision threshold.

A variation of an accumulation account has also recently been proposed in the domain of metamemory (the *self-consistency model*; Koriat, 2012; Koriat & Adiv, 2016). In much the same way subjects in SDT models are thought to compute the 'statistics' of their perceptual choice, the self-consistency model proposes that subjects internally replicate the decision several times, drawing samples of their decision and computing the internal reliability of their choice. Samples are theorized to accumulate over time until a pre-set number of samples converge on the same answer (akin to the decision threshold in other accumulator models).

Confidence in these models can potentially come from two places. First, if time is unlimited, then the length of time it takes to reach a decision directly reflects our confidence; fast decisions are ones where one option was very likely and we should be highly confident (Kiani & Shadlen, 2009). Second, if time is limited, then the amount of accumulated evidence (e.g., how far past the midpoint team A has pulled the rope) directly reflects our confidence (Baranski & Petrusic, 1998). There is also some heterogeneity between models about when in the accumulation process confidence is computed, and what information is available to the accumulator to influence the tug-of-war. In some models, the computation of confidence occurs at the time of the decision (Gigerenzer et al., 1991), while other models allow for post-decisional

information to continue accumulating and thereby modify a later confidence computation (Baranski & Petrusic, 1998; Pleskac & Busemeyer, 2010). Within most perceptual models, 'information' that accumulates is thought to be directly and entirely based on the noise of the Type 1 representation, as in SDT (i.e., the noisier the representation, the less predictable the tug-of-war will be). However, some models, like the self-consistency model of metamemory, explicitly allow other influences on the Type 1 decision (Koriat & Adiv, 2016), though even these models argue that the same information is used to make the Type 1 decision as the subsequent Type 2 decision.

Because confidence evidence is thought to come from the same source as decision evidence, the Accumulator account predicts that Type 1 and Type 2 decisions should be tightly linked, just as SDT models do. This link may not be a perfect 1:1 correspondence if a post-decisional accumulation period is included in the model, as this would affect the confidence judgment but not the decision and thereby dissociate the two (Pleskac & Busemeyer, 2010). However, much like SDT models, an Accumulator account predicts that this tight link between Type 1 and 2 evidence should also reveal itself in a lack of correlation between Type 2 decisions dependent on dissociated Type 1 evidence. Also, like SDT, accumulator models predict that developmental change, if any, is due to changes in decision noise or the position of the decision threshold, as there is an implicit assumption that this process should be in place throughout the lifespan.

Summary. Across this first family of accounts, we can see that confidence is critically conceptualized as a direct output from the same noisy information used to make the decision. These Direct accounts therefore all make three key predictions. First, if we can isolate the noise of Type 1 abilities and the noise of Type 2 abilities separately from the placement of decision thresholds for each, we should find near-perfect overlap between the two levels of noise, a prediction I test in Chapter 2. Second, because Type 2 noise is tightly coupled to Type 1 noise, there should be no strong relationship between Type 2 abilities across distinct tasks that rely on distinct and uncorrelated Type 1 representations (e.g., emotion vs. number decisions), a prediction I test in Chapter 3. And third, these patterns should hold across development, and the only potential source for developmental change is from the improvement of Type 1 abilities or in the placement of the decision threshold, a prediction I test in Chapter 2. A summary of these predictions is available in Figure 1.1.

Figure 1.1

Predictions Made by the Two Accounts.

	Type 1 and Type 2 Noise	Development of Metacognition	Metacognition for Independent Type 1 Tasks	Self and Other Confidence
Direct Account	• Highly correlated	• In Type 1 noise • In response biases	• Unrelated • In Type 1 units	• Unrelated
Inferential Account	• Related, but dissociable	• In Type 1 noise • In response biases • In other sources	• Correlated • In common units *(Bayesian)*	• Highly correlated *(Mindreading)*

Inferential Accounts

Critics of the Direct accounts have noted that confidence does not seem to perfectly derive from the quality of underlying representations (Koriat & Levy-Sadot, 2001). Within the study of metamemory, several studies have noted the presence of metamemory illusions, where participants' confidence can be manipulated by varying properties of the task that should be irrelevant. For example, adults' confidence is higher for items written in clear font, or when words are primed with related concepts (for review, see Alter & Oppenheimer, 2009). To account for the influence of these external factors, several proposals for how confidence is computed argue that confidence is the result of an inferential process that combines information from several sources (Koriat, 1993; Koriat & Levy-Sadot, 2001; Nelson & Narens, 1990). In this family of accounts, subjects are thought to compute confidence not only from the noise in their representations (if at all), but rather from the presence of other cues that signal likely accuracy.

Cue-Based Accounts. Many behavioural and cognitive cues to confidence (or sometimes *heuristics*) have been identified, largely in the context of memory judgments. For instance, decision confidence is said to be based in part on the presence of *accessibility* cues like the speed of access to a memory or the number of alternative choices a participant could think of (Koriat, 1993). While many individual cues have been identified over many years (e.g., Alter & Oppenheimer, 2009; Metcalfe et al., 1993), this account broadly assumes that subjects have little to no direct access to the content of their decisions and instead uses both external cues (like reaction time) and internal cues (like the availability of alternatives, which cannot be directly observed by others to infer the strength of one's knowledge).

Cue-Based accounts rest on the assumption that confidence is not a direct read-out of decision noise, and therefore strongly predict that Type 1 and Type 2 reasoning are dissociable – being skilled at remembering word pairs (a common Type 1 task) does not mean that you are skilled at keeping track of decision time (a commonly proposed Type 2 cue). These accounts also hint towards confidence computations occurring similarly between different domains[2] (e.g., observers good at keeping track of decision time for memory judgements should be equally good at keep track of time for perceptual judgements), predicting a strong relation between confidence reasoning across domains. As with Direct accounts, most Cue-Based accounts do not predict development specific to the process of computing confidence. If development occurs, this is attributed to changes in children's attention to certain cues with experience (e.g., noticing over time that reaction time predicts accuracy), rather than their ability to use cues to compute confidence (Koriat et al., 2009). Thus, it is challenging to form specific predictions about development *a priori,* except perhaps that if a cue is shown to be relevant for one kind of confidence judgment, it should be relevant for all confidence judgments (e.g., if a child attends to reaction time to cue confidence in a counting task, they might also attend to reaction time to cue confidence in a memory task).

The Mind-Reading Account. As a social species, humans must frequently reason about mental states when interacting with others, including deciphering the knowledge that others have so we can communicate effectively (Baer & Friedman, 2018), or thinking about which friend might be more knowledgeable about calculus and be able to tutor us (Koenig & Harris, 2005b). Given that metacognition involves monitoring one's own mental states, another account for how we understand our own minds is that we turn our abilities for reading *others'* minds inward upon ourselves (Carruthers, 2009), tracking our own observable behaviour. For example, in the same way we might notice that a friend tends to be accurate whenever they answer quickly, we might predict our own accuracy by tracking how quickly we answered.

This account therefore makes two key predictions. First, the need for an inferential process suggests that there is limited (if any) access to internal representations (Carruthers, 2009), predicting a dissociation between Type 1 and Type 2 abilities. That is, just because a

[2] Note that a lack of similarities between decisions in different domains does not falsify this account, it could simply mean that each domain has access to different cues (e.g., memory might be affected by the number of available alternatives in a way that perception would not).

subject is skilled at detecting differences in quantities does not mean they can interpret their own behavioural cues effectively. Second, because of the limited access to internal representations, we would expect that metacognitive reasoning about confidence should operate identically across all domains of cognition.

Additionally, this account makes one very specific third prediction: metacognitive ability and mind-reading ability, more commonly referred to as *Theory of Mind* (Baron-Cohen et al., 1985), are one in the same, which should lead to strong correlations between the two abilities, a prediction I test in Chapter 4. Moreover, most proponents of this account specifically predict that the capacity for reasoning about others' mental states should emerge *prior* to metacognition, meaning that metacognitive abilities should never exist without there also being mind-reading abilities (though mind-reading could exist without metacognition; Carruthers, 2009). Given that it is well documented that social cognition changes in development (e.g., Wellman & Liu, 2004), this also points to a strong developmental prediction: age-related change in Theory of Mind should be tightly related to change in metacognition.

The developmental prediction about Theory of Mind does not have strong empirical support, however. When initially proposed, there was only evidence of mind-reading capability only in children older than 3 and not in children with autism spectrum disorder or in other species (Baron-Cohen et al., 1985; Premack & Woodruff, 1978; Wimmer & Perner, 1983), and evidence of metacognition only in children aged 7-8 and older (Flavell, 1979), supporting the claim that mind-reading developed first. However, as discussed in the section "Do Children Have a Sense of Confidence?", many non-verbal tasks demonstrate the presence of metacognitive reasoning in young infants, children with autism spectrum disorder, and several other species (Elmose & Happé, 2014; Goupil & Kouider, 2019; Kepecs & Mainen, 2012).[3] Thus, simply arguing that Theory of Mind appears earlier in developmental or evolutionary time does not conclusively support a tight link between the two. Moreover, early evidence stating that the two abilities were related largely relied on tasks that explicitly asked children to report their own or another's mental state (e.g., asking if the child 'knew' what an object's secret property was before they observed it; e.g., Gopnik & Astington, 1988). However, as already discussed,

[3] With several creative new methods, Theory of Mind is now thought to emerge much younger, by at least the second year (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2017) and potentially even in apes (Krupenye et al., 2016), though there is considerable controversy around this (Kulke et al., 2018; Kulke & Rakoczy, 2018). As such, it is hard to make clear statements about the dissociations between the two abilities.

this confounds linguistic competence (and potentially other developing abilities; see Proust, 2012), while leaving it unclear whether the ability to reason about the self and others is related beyond the shared linguistic markers.

The Bayesian Account. An influential account in modern cognitive science is that of Bayesian rationality, which at its most basic level argues that we form beliefs by combining new evidence with our existing knowledge (Glymour, 2003; Gopnik & Bonawitz, 2015; Xu & Tenenbaum, 2007). Much like SDT, Bayesian models specifically assume that our representations are noisy probability distributions. However, the Bayesian account takes this one step further and critically conceptualizes confidence as *the probability of a decision being correct*. This critical addition therefore puts all decisions in common units (probability of being correct) and not in the original Type 1 units (Meyniel et al., 2015; Pouget et al., 2016). If we return to the analogy of statistical analyses, in which certainty is the standard deviation and the SDT characterization of confidence is an effect size, the Bayesian conceptualization of confidence is like a standardized effect size that can be meaningfully compared between different tests (e.g., an odds ratio). The consequence is that this probability can then be rationally combined with other probabilities according to Bayes' rule to form a single estimate of confidence (e.g., drawing repeated samples over time, or combining information from multiple cues; Meyniel et al., 2015).

Conceptualizing confidence as a probability means that Bayesian accounts generally predict that there will be a link between Type 1 and Type 2 decisions, but that the conversion to probability of a decision could introduce noise unique to the Type 2 decision, dissociating the two (Meyniel et al., 2015). Moreover, as an Inferential account the Bayesian account does not *require* that confidence be calculated from Type 1 noise – the probability of an answer being correct could be determined from other cues like decision time as well, as pointed out by the Cue-Based account. In fact, the probability of a given decision could even be estimated separately from different cues and then integrated to form a more robust feeling of confidence (Barthelmé & Mamassian, 2010). For instance, if you are watching a firework show, you can use your prior knowledge about the light travelling faster than sound to increase your confidence that the source of a loud bang was from the explosion and not another source (though your sensory systems without this knowledge would interpret the light and sound as separate events). This could, under some circumstances, lead to Type 2 decisions that do not use Type 1 information at

all, fully dissociating the two abilities (say, if you chose to ignore sounds entirely because you know they travel slower than light). In other words, the Bayesian account can explain both a tight link and a dissociation between Type 1 and Type 2 abilities.

Nevertheless, there is one clear prediction that arises from the Bayesian account. Because confidence is represented in units indifferent to the Type 1 decision (or what we might consider *domain-general* or *domain-neutral* units), we should strongly expect that Type 2 abilities are related to one another. More specifically, we would expect that subjects should compare and integrate confidence across independent Type 1 tasks with ease, as the unit of confidence is identical (De Gardelle & Mamassian, 2014). And finally, because the Bayesian account assumes that reasoning about probability is fundamental to decision-making, it predicts that even children should use domain-general units to reason about confidence (Vo et al., 2014).

Summary. Within these three groups of Inferential accounts, we see a strong common theme that the human mind is critically integrative, using information from multiple sources to inform decisions. While each specification of an Inferential account makes different predictions about how tightly linked cognitive and metacognitive abilities are, they all importantly allow for there to be information in the Type 2 confidence decision that was not part of the Type 1 decision. Because of this, Inferential accounts predict that confidence reasoning is related for multiple independent Type 1 decisions. In particular, the Bayesian account predicts that all confidence judgments are represented as the probability of being correct, and should therefore be compared and combined flexibly, even in childhood. A summary of these predictions is available in Figure 1.1.

Disentangling the Accounts of Metacognition

Each of these accounts describes confidence in the same general terms (an estimate of the strength of a percept, memory, or decision), but they differ critically in the sources that lead to this estimate and precisely how such an estimate is calculated. Because of these differences, each account makes specific predictions about two properties of metacognition: (1) how tightly bound confidence is to the cognitive processes it evaluates (i.e., the link between Type 1 and Type 2 decisions), and (2) how broadly is confidence represented in the mind (e.g., whether different Type 2 decisions relate to one another, or to other-oriented mind-reading abilities). In each of the following research chapters, I investigate these questions using development as a lens for testing

between these theories and in turn building a better picture of metacognition in childhood as a driver of rational thought.

In Chapter 2, I pit the theoretical accounts against one another by investigating whether there is any developmental change associated with confidence judgments once response biases are accounted for. For instance, the Direct accounts strongly predict that children's Type 1 accuracy (an index of Type 1 perceptual imprecision) will entirely explain any variability in children's metacognitive performance. In contrast, the Inferential accounts predict development that cannot be explained by Type 1 accuracy. To accomplish this, I introduce a new paradigm for the study of metacognition in children that does not require children to provide a confidence judgment on any scale (which could facilitate biases), but rather asks children to relatively compare two states of confidence. This measure, in combination with an assessment of children's accuracy on the Type 1 task, allows me to test between competing predictions from the theoretical accounts.

In Chapter 3, I conduct a further test between these accounts by looking for signatures of domain-generality in confidence reasoning. The Direct accounts predict domain-specificity in childhood, that metacognition should operate independently for every cognitive process it evaluates. The Inferential accounts, however, predict commonality in metacognitive reasoning. Therefore, I adapt the paradigm introduced in Chapter 2 to collect metacognitive assessments of three independent domains and test whether children's metacognitive abilities are shared across domains or operate independently. Additionally, the Bayesian account makes the strong prediction that confidence is a common currency between independent tasks, and so I also test whether children can compare confidence between tasks and whether the units of confidence are interchangeable.

Finally, in Chapter 4 I focus in on the link between evaluating one's own abilities and evaluating the abilities of others to determine *how* broadly confidence is computed. Here, the Mind-Reading account (and to some degree the other Inferential accounts) make a very strong prediction that the two should be nearly identical, and the Direct accounts predict that they should be entirely unrelated. To this end, I adapt a task designed to examine whether children can detect relative differences in others' accuracy to match the relative confidence task used in Chapters 2 and 3. The presence or absence of a strong correlation between the two tasks will

further disentangle the competing accounts, and also point to how metacognitive reasoning can be used by children to rationally evaluate information.

Chapter 2: Accounting for Developmental Change in Metacognition

As children age, their metacognitive abilities improve on most measurements (Ghetti et al., 2013). For instance, children at age 7 do not modify study time in response to difficulty while at age 9 they do (Lockl & Schneider, 2004), and children at age 5 show less consistency in their use of high and low bets in response to uncertainty than at age 8 (Vo et al., 2014). Direct accounts of confidence propose two very straightforward mechanisms for this developmental change: changes in the placement of the threshold between high and low confidence and the reduction of noise in Type 1 abilities as they develop. If confidence judgments are a direct readout of Type 1 noise, then removing these two sources of variability should entirely remove the age-related variability of Type 2 confidence decisions. In contrast, Inferential accounts predict that there could be remaining developmental variability (in the inferential process, or in attention to different cues). As a first step to teasing apart these families of accounts, we therefore need to measure children's metacognitive abilities separately from their response biases that affect the placement of their confidence threshold and separately from their Type 1 abilities.

Accounting for Response Biases

In the developmental literature, many studies have found that children are overconfident, reporting that they have known facts all along or frequently selecting the often happy-faced 'high confidence' response on a scale (Paulus et al., 2013; Taylor et al., 1994; van Loon, de Bruin, et al., 2017). While interesting in its own right,[4] children's frequent overconfidence is a *bias*, a systematic tendency to misreport their experienced confidence. Bias is potentially informative about how we *use* confidence judgments, but could artificially induce noise in our measurement of the underlying *signal* of confidence, leading to an apparent dissociation between confidence and accuracy. Accordingly, one critical observation is that overconfidence decreases with age (Lockhart et al., 2017; van Loon, de Bruin, et al., 2017; Vo et al., 2014), suggesting that developmental effects may not only be driven by change in accuracy, but by change in bias (Salles et al., 2016).

[4] Though not the focus of this dissertation, it is worth noting that such overconfidence biases may actually be adaptive for young learners, pushing them to attempt tasks slightly outside their current ability level as a kind of self-scaffolding or protecting them from harshly evaluating their failures (Bjorklund & Bering, 2002; Kidd et al., 2012; Lockhart et al., 2017), prompting the possibility that these biases are a critical part of human cognition.

To illustrate how bias could artificially induce variability in metacognitive performance, think of a task of perceptual brightness: you must identify whether a given shade of grey is dark or light. If you have a perfect sense of brightness, you will indicate 'light' for every shade that is above 50% brightness and 'dark' for every shade that is below. Now, imagine that you are completing this task while wearing sunglasses. Suddenly, everything under 75% brightness appears 'dark' to you! Your ability to tell light from dark is not altered (the greys themselves are still the same, and you are still you, just much cooler), but your performance is suddenly very poor because of your sunglasses. Analogously, participants with strong biases (with sunglasses) may appear to have poor metacognitive abilities because they reserve the use of low or high confidence for very extreme cases, but this does not mean that their ability to reason about their confidence is any worse than other participants (without sunglasses; Fleming & Lau, 2014; Maniscalco & Lau, 2012; Nelson, 1984).

The distinction between ability and bias is the basis for SDT (Green & Swets, 1966). Based on classic psychophysical studies, SDT points out that whenever we report an experience from a continuous dimension (like brightness or confidence), that report is fundamentally comprised of two components: our *sensitivity* to that dimension and our *bias* towards one endpoint over the other. Sensitivity, sometimes referred to as *resolution* or *precision*, reflects our ability to discriminate points on the continuous dimension, like '12% certain' from '50% certain' or '13% certain'. Bias, sometimes referred to as a *criterion* or *calibration*, reflects the threshold we use to divide the continuous dimension into discrete categories, like dividing low from high confidence at '50% certain'. The challenge illustrated by the sunglasses example is that the participant sets this threshold depending on the biases that they possess, and the researcher must account for such biases in order to truly understand that participant's sensitivity.

Because the issue of biased criterion-setting is not new to the study of metacognition, there have been many proposed tools for researchers to correct this bias. The first is to rely on statistical analyses that isolate confidence sensitivity separately from confidence bias. Early methods relied on non-parametric correlations between accuracy and reported confidence, such as phi or gamma (Fleming & Lau, 2014; Galvin et al., 2003; Maniscalco & Lau, 2012; Nelson, 1984), but these measures do not actually disentangle sensitivity and bias – they are simply less influenced by bias than reporting a participant's average confidence rating. More recently, researchers turn to statistical methods based on the logic of SDT that model both sensitivity and

bias based on the participant's metacognitive accuracy (e.g., indicating high confidence on accurate trials), including calculating the area under the Type-2 Receiver Operating Characteristic curve (AUROC2; Galvin et al., 2003; Vo et al., 2014), or modelling the participant's expected sensitivity based on their confidence judgments (meta-d'; Fleming, 2017; Maniscalco & Lau, 2012, 2014). Simulations of these measures revealed that they successfully reduce the influence of bias on estimates of sensitivity (Fleming, 2017; Maniscalco & Lau, 2012, 2014).

However, the solutions developed to combat this issue have not been readily adopted by developmental metacognition researchers. Only two studies to date in children under age 8 have used SDT-based corrections, one finding that there was still development once bias was accounted for (Vo et al., 2014) and the other finding no development (Salles et al., 2016). Part of the challenge for developmental researchers in particular is that statistical corrections like AUROC or meta-d' require a large number of trials (ideally over 100), and that participants have considerable variability in their answers (Fleming, 2017). Given that many studies with children are necessarily limited by their attention span, it can be difficult if not impossible to get more than 50 trials of one task with young participants.

Luckily, there is a second tool to eliminate the influence of bias without the use of statistical measures. The solution is to control for bias experimentally by using a *relative* assessment of metacognition, in which participants must directly compare the strength of two states of confidence (Barthelmé & Mamassian, 2009; Butterfield et al., 1988; Lipowski et al., 2013). Consider that when you look at the world wearing sunglasses, even though everything looks darker than normal, you do not lose the ability to tell apart lighter areas from darker ones. That is, your bias to see the world as darker overall does not influence the resolution with which you can tell dark from light in the world. This observation leads us to an experimental paradigm, in which subjects see two alternatives and must select which best fits a given description, like which is darker or which elicits a stronger sense of confidence. These *two-alternative forced choice* (2AFC) tasks have been widely adopted in the broader psychophysical literature not only because they bypass the influence of bias entirely, but because they allow for greater experimental control in quantifying sensitivity (e.g., presenting both difficult comparisons and easy comparisons to identify precisely what the 'just noticeable difference' on a continuum is; Weber, 1978). 2AFC task have also been widely adopted in developmental psychology because

they can easily be used with little to no language demands (e.g., asking children "which group has more" instead of asking children to estimate a number of objects; Halberda & Feigenson, 2008). However, with a few notable exceptions (Barthelmé & Mamassian, 2009; Butterfield et al., 1988; Lipowski et al., 2013), relative measures of metacognition have been largely underused both in the study of metacognition and specifically in research on children's metacognition.

Accounting for Type 1 Development

The use of a relative assessment of confidence helps eliminate the known development of overconfidence biases, but it does not yet fully test the prediction of the Direct accounts that Type 1 noise is the source of Type 2 metacognitive ability (Maniscalco & Lau, 2012). To illustrate this prediction, consider a task in which you must choose the darker of two shades of grey and then subsequently report your confidence in that decision. These decisions will 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 confidence decision would *appear* imprecise because it also only has two states: impossible or trivially easy. Variability in Type 1 abilities (which we would strongly expect in many developmental samples) might therefore be partially or completely responsible for the appearance of improving metacognitive ability.

There are generally three approaches to combatting this concern and detecting whether there is any unique variance in Type 2 abilities that cannot be accounted for by Type 1 abilities. First, the experiment can be set up in such a way that every participant achieves the same overall accuracy on the Type 1 task, either by modifying the stimuli at the group level to account for expected developmental change (e.g., Selmeczy & Ghetti, 2019), or by adjusting the difficulty of the questions every few trials to adapt to the participant's current performance (e.g., Vo et al., 2014). These latter strategies can be somewhat problematic, though, because children's performance on any given trial is influenced by the difficulty of the preceding trials (Odic et al., 2014; Wang et al., 2016), and metacognitive judgments specifically appear to be strongly based on the accuracy of preceding trials (Martí et al., 2018; Rahnev et al., 2015).

A second method that does not run into such issues is to statistically control for individual differences in Type 1 performance (e.g., through a partial correlation or in a regression). For example, Maniscalco and Lau (2012) computed measures of the sensitivity of both Type 1 and Type 2 abilities using the same units (*d'*, a standard tool in SDT; Green & Swets, 1988), and formed a ratio of the two which they called *metacognitive efficiency*. Overall, participants had worse Type 2 sensitivity than Type 1 sensitivity, indicating that the confidence calculation had less information available to it than the Type 1 decision.[5] From this evidence, it appears as though Type 1 noise and bias are not the sole contributors to metacognitive abilities.

If true that there is additional information beyond Type 1 noise and bias to make confidence judgments, as predicted by Inferential accounts, we would expect this information to have meaningful variability (e.g., a correlation with the cognitive abilities linked to tracking cues). To test this prediction, we can leverage development as a source of meaningful variability across many potential cues, as many skills relevant to tracking potential cues develop (e.g., reasoning about time, evaluating facial expressions; Odic, 2018; Thompson & Lagattuta, 2006). Moreover, we might expect that over time, children become better able to appropriately weigh Type 1 noise and other cues (Vo et al., 2014). We would therefore expect that if we can effectively control for both bias and Type 1 sensitivity, children's confidence judgments will still correlate with age.

To accomplish this, we use children's number sense as a case study. The Approximate Number System (ANS) is the theorized evolutionarily-adapted system for representing numerical information that guides our earliest intuitions about number (Dehaene, 2011; Halberda et al., 2012; Odic & Starr, 2018; c.f. Gebuis & Reynvoet, 2012; Szűcs et al., 2013). It is present in newborn infants (Izard et al., 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

[5] Note that it is theoretically possible for Type 2 abilities to have *more* information available to them, say if the accumulation process reveals more evidence after the Type 1 decision was made, or if certain cues are available to the confidence decision.

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. In fact, asking children to determine the larger of two quantities is an example of the relative 2AFC task described earlier that can isolate sensitivity from bias, giving us a simple way to isolate Type 1 sensitivity.

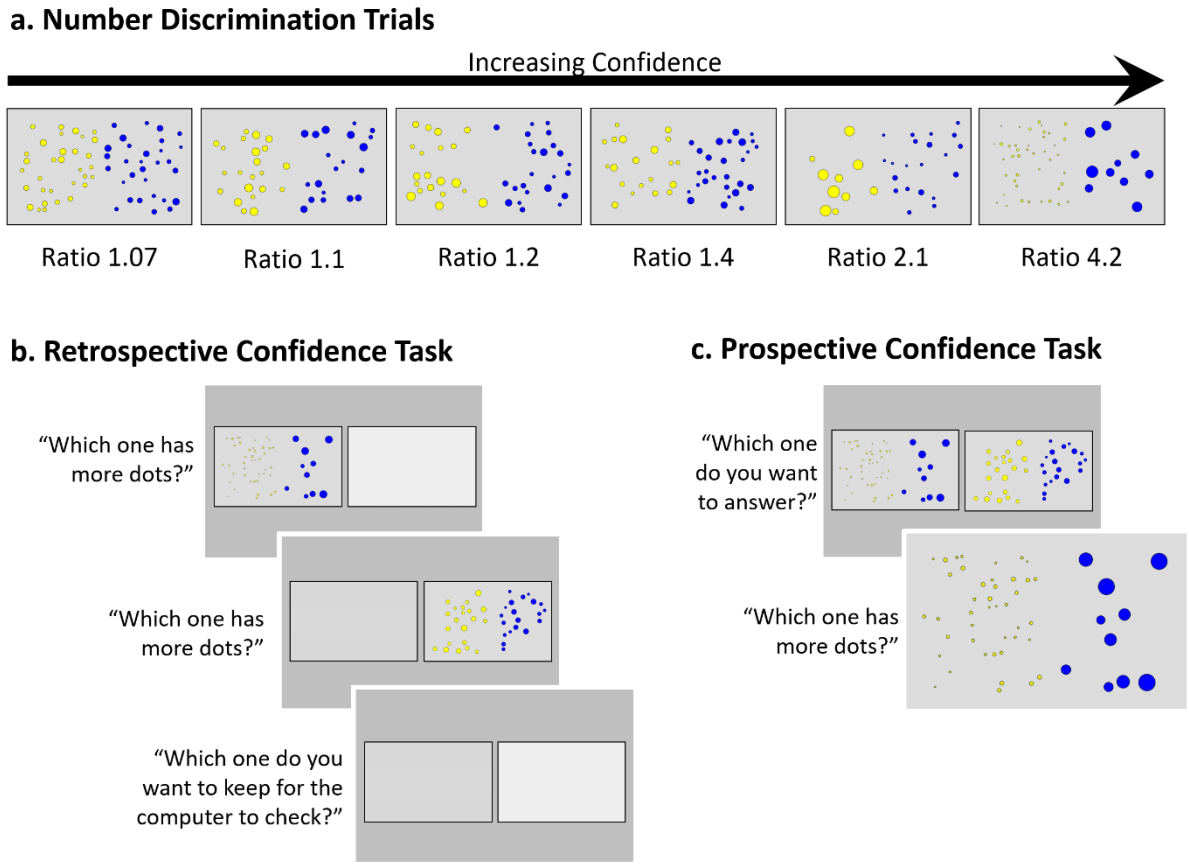
Importantly, recent theoretical and empirical work has shown that the ANS provides us with both an approximate sense of number *and* a sense of our confidence in that estimate. For example, if you were asked to estimate the number of words on this page, your ANS would provide you with both the most likely number but also a sense of how confident you should be in that value (Halberda & Odic, 2014; Vo et al., 2014). Young children can also reason about their confidence in simple ANS decisions: after completing a number discrimination trial (deciding whether there are more dots on the left or right side in Figure 2.1), 5-8 year-old children can also indicate whether they believe that they answered the trial correctly or incorrectly by placing "high" or "low" bets (Vo et al., 2014). Because even young children can reason about numerical quantities, there is development in numerical precision, and children experience confidence in numerical judgments, the ANS is an ideal testing ground for the accounts of developmental change in metacognition.

Then, we need a method of measuring metacognition that both controls for children's response bias and that can allow us to measure the precision of the ANS independently from confidence decisions. To isolate metacognitive sensitivity separately from bias, we therefore use a relative confidence task, in which children must select which of two answers they are most certain of. Using a relative task greatly reduces alternative explanations like developing language understanding and does not require as many trials as tasks that use a statistical correction for bias, making it ideal for use with children. Then, we statistically control for ANS abilities,[6] examining whether there is any remaining developmental change in metacognitive abilities. If there is continued development in metacognitive abilities when controlling for both response biases and ANS abilities, we would have evidence that there is meaningful change specifically in metacognitive abilities.

[6] At present, the metacognitive efficiency measure is only available for tasks where participants select one of two answers and then rate their confidence in that choice, and thus cannot be used with the relative paradigm.

Figure 2.1

Sample Stimuli Used in Study 1.



Note. Section a depicts sample number discrimination trials in which children indicate which colour has more dots. Section b depicts the Retrospective Confidence 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 Prospective Confidence Task, in which children first answer the confidence question by selecting the trial they most expect to get correct, then answer only that question.

Study 1

Methods

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 2.1 for distributions by age). All children were tested in a quiet space in their schools or daycares in Vancouver, BC. No additional demographic information was collected, though most children

were middle- to upper-middle class and largely from White or East/South-East Asian backgrounds. All children spoke enough English to carry a simple conversation with the experimenter.

Materials and Procedures. Tasks were presented on an 11.3" Apple Macbook 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.

Table 2.1

Sample Sizes, Means, and Tests Against Chance for the Number Task, and Confidence Tasks in Study 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
Retrospective Confidence 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
Prospective Confidence 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

Stimuli throughout the experiment consisted of trials from a number discrimination task used widely in the literature on the ANS that uses a 2AFC method to eliminate the influence of Type 1 response biases (e.g., over-or under-estimation; Halberda et al., 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 Figure 2.1). 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 Figure 2.1a depicts a ratio of 4.2 (42 yellow dots and 10 blue dots), which elicits a high degree of confidence, and which most children in this age range would get correct (Odic, 2018). In contrast, the first image in Figure 2.1a depicts a ratio of 1.07, which elicits a much lower degree of confidence, and which very few children in this age range answer correctly above chance rates. 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 confidence sensitivity, we designed two versions of the relative confidence 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 gave them higher confidence. 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). We will refer to the first version as the Retrospective Confidence task because the confidence judgment is made *after* the perceptual decisions, and the second version as the Prospective Confidence task because the confidence judgment is made *before* the perceptual decision.

Each version of the relative confidence task has its own strengths and limitations. The Retrospective task allows us to simultaneously collect confidence and perceptual judgments for all trials, while the Prospective task does not, because children only answer the one question they indicate as high confidence. However, the Retrospective version potentially places additional cognitive and motivational demands on children that the Prospective version does not. Completing the Retrospective task requires that children hold in memory their two states of confidence from the preceding perceptual decisions, overcome cognitive fatigue to report on their confidence 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 confidence - 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.

Retrospective Confidence Task. In this task, children were shown two gray occluders – one on the left side of the screen and one on the right (Figure 2.1b). 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 were 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 colour 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 confidence 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 confidence judgments in children as young as 3 years (Hembacher & Ghetti, 2014; Vo et al., 2014). As in Barthelmé & 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 confidence. 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 *confidence* choice was correct or not (see Smith, et al., 2008).

Critically, the trials were paired such that one always displayed a larger (i.e., higher confidence) ratio than the other. We expected that children would choose the answer they felt was more certain. To assess individual differences in sensitivity to confidence, 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). The difference in difficulty between the two trials becomes harder to detect as the metaratio approaches 1.0, much like the difficulty in telling apart two quantities becomes harder to detect as the ratio approaches 1.0. 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 confidence 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 et al., 2016), and our choice of using division does not impact any of our results.

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 confidence questions – i.e., which trial they were more certain of. This task took, on average, 5.6 minutes for children to complete.

Prospective Confidence Task. The stimuli for this version were identical to the Retrospective version: the identical 60 number discrimination trials were used in exactly the same pairings as in the Retrospective version to limit the differences between the tasks. However, in the Prospective version, both number discrimination trials were visible side-by-side on the screen at the beginning of each trial (Figure 2.1c). Rather than answering each question and then retrospectively evaluating their confidence, children were instead asked "Which one do

you want to do?" (capitalizing on children's desire to maximize their success). 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 confidence prospectively and chose a trial to complete based on their perceived higher confidence. Children were given as long as they needed to answer both the confidence and number discrimination questions, though they were discouraged from counting in the same way as in the Retrospective 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 confidence 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. This task took, on average, 3.6 minutes for children to complete.

Results

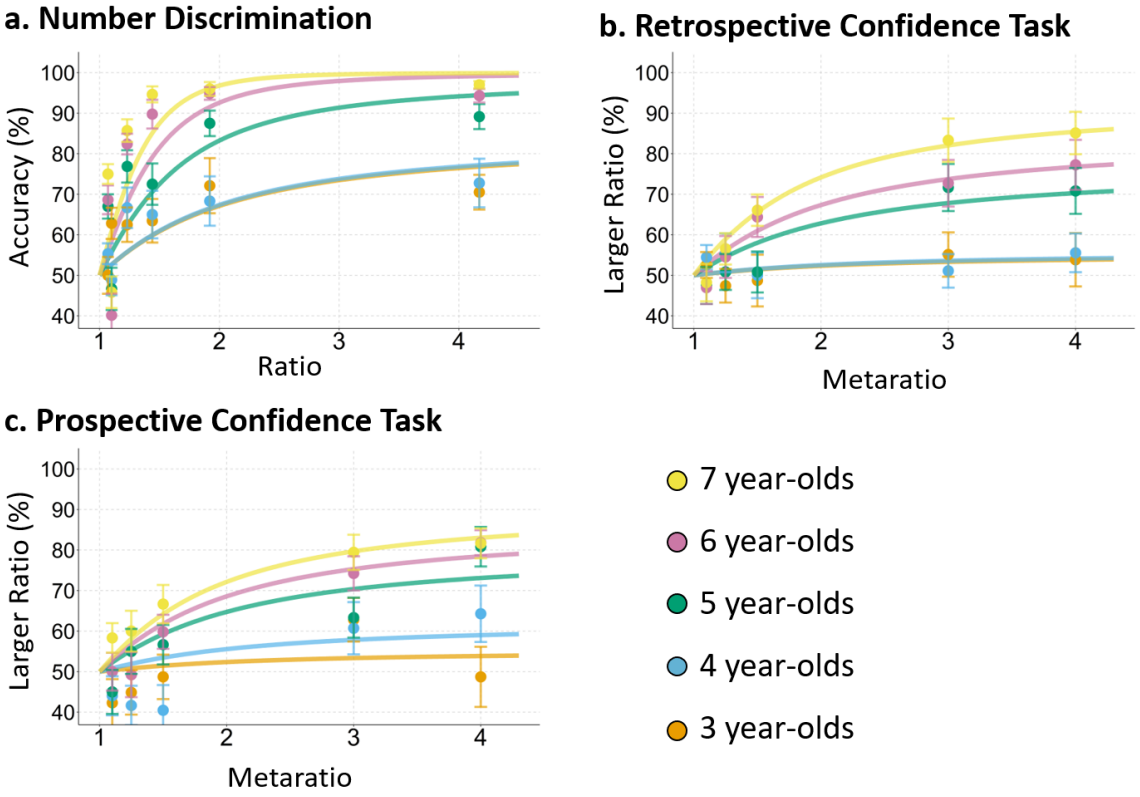
We found no effects of gender in our analyses, so all results reported hereafter collapse across gender. Children were generally more accurate on the Prospective Confidence task if they completed it first, likely because the longer Retrospective task was more fatiguing, $F(1, 92) = 4.73$, $p = .032$,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

Number Discrimination. Children's average accuracy on the number discrimination trials within the Retrospective 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 Figure 2.2, 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 confidence, as their accuracy improved with larger ratios.

Figure 2.2

Accuracy by Ratio on the Number Discrimination Trials, and by Metaratio on Confidence Trials in Study 1.



Note. Error bars represent 1 SE, and curves are estimated using a standard psychophysical model (see Odic, 2018).

Retrospective Confidence Task. Because each trial consisted of a smaller and a larger ratio, we expected that children who attended to and compared two states of confidence 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 2.1 for means and tests against chance)[7]. 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 Prospective task to determine which question to answer. Using 1000 random splits, the Spearman-Brown corrected reliability estimate for how often children chose the more certain ratio was .72, 95% CI [0.63, 0.79] (Parsons, 2020).

As a further examination of whether children chose trials based on their confidence, 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 confidence 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 most 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 confidence.

Next, we turn to the central question of interest: which factors predict the development of children's confidence? We found a strong correlation between children's choices on the Retrospective Confidence task and age, $r(96) = .40$; $p < .001$. Since the relative confidence task eliminates response biases, this result suggests that children's ANS confidence sensitivity develops independently of their bias. We found this same result when examining the correlation between age and the ANS accuracy on chosen vs. discarded trials, $F(4, 93) = 6.44$, $p < .001$; $\eta_p^2 = .22$.

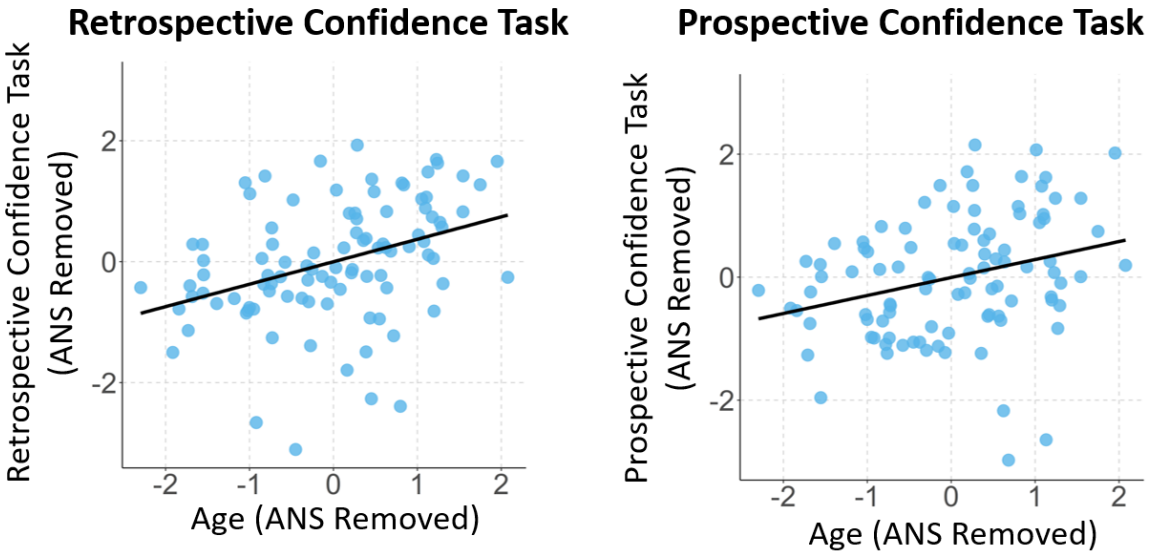
But, could this age-related improvement simply be due to children's improving ANS precision, as predicted by Direct accounts? We found a trending correlation between children's ANS discrimination accuracy and their choice on the confidence task, $r(96) = .19$, $p = .066$, suggesting that the ANS contributes some variance to children's performance on the confidence task. However, adding ANS discrimination ability to a linear regression between confidence and age did not improve the model predicting confidence 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 Figure 2.3), suggesting that there are age-related improvements in confidence sensitivity *independent* of the underlying improvements in ANS representations themselves. But,

[7] 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.

do we observe identical results in the Prospective Confidence task, in which children have to evaluate their confidence prospectively rather than retrospectively?

Figure 2.3

Partial Correlations Between Confidence Accuracy and Age, Controlling for Number Discrimination (ANS) Accuracy in Study 1.



Prospective Confidence Task. As in the Retrospective Confidence task, we found that children ages 5, 6, and 7 in the Prospective Confidence task chose the more certain ratio more than 50% of the time (see Table 2.1 for means and tests against chance)[8], and age correlated with the Prospective accuracy, $r(97) = .47$, $p < .001$, suggesting again that response bias is not the only factor responsible for the development of ANS confidence. Spearman-Brown corrected reliability was .72, 95% CI [.64, .80]. We also found a small age-related difference between the two Confidence tasks: as can be seen in Figure 2.2, even 4-year-olds selected 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 2.1). Therefore, it is possible that young children's ANS confidence is so noisy and

[8] As in the Retrospective 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.

imprecise that they cannot reliably tell apart the metaratios we presented, but that they might succeed if given easier metaratios.

Replicating the Retrospective 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 Figure 2.3, suggesting that the development of sensitivity to confidence is not entirely driven by improvements in the underlying perceptual representations themselves, even in a prospective task with reduced cognitive demands (counter to the predictions of the Direct accounts).

Correlations Between the Tasks. Because the Prospective and Retrospective Confidence versions differed in several ways, we performed two additional comparisons between tasks to confirm that both versions were measuring the same underlying ability. First, confidence 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 confidence in (trials they chose to answer in the Prospective version, and trials they chose to keep in the Retrospective version) were nearly identical (Prospective: $M = 85\%$, $SD = 11\%$, Retrospective: $M = 84\%$, $SD = 14\%$), $t(96) = 0.29$, $p = .771$. In fact, these two accuracies correlated even when controlling for age, $r(94) = .37$, $p < .001$, suggesting that children were trying to choose questions in both versions that maximized their chance of success.

Together, these results show that the Prospective and Retrospective tasks both tapped into children's confidence and that, furthermore, the development of children's confidence sensitivity occurs independently of response biases and individual and developmental differences in ANS acuity itself.

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 confidence 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 confidence 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 Retrospective 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 Figure 2.2) and the Prospective task (Accuracy: $F(4, 376) = 28.41$, $p < .001$, $\eta_p^2 = .23$; Speed: $F(2.67, 261.24) = 4.01$, $p = .011$, $\eta_p^2 = .04$; Figure 2.2). Children's age (as a covariate) interacted with accuracy by metaratio in the Retrospective 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 Prospective 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 confidence are themselves subject to internal noise and are consistent with Weber's law (Weber, 1978).

Exploratory Analysis of the "Opposite Strategy". As noted above, we found that a small subset of children in both the Prospective and Retrospective tasks consistently chose the trial they were *less* certain of ($n = 12$ in the Retrospective Confidence task and $n = 11$ in the Prospective Confidence 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 Odic et al., 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 behaviour in both the tasks.

In the analyses reported above, we left all 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 behaviour clearly indicates an ability to differentiate their two states of confidence. 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 Retrospective task, $r(85) = .52$; $p < .001$, and the Prospective task, $r(87) = .53$; $p < .001$. Both of these remained significant even when controlling for individual differences in Number Discrimination accuracy, Retrospective: $r(84) = .46$, $p < .001$; Prospective: $r(84) = .43$, $p < .001$. We also 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 Retrospective task, $r(96) = .49$; $p < .001$, and Prospective task, $r(97) = .50$; $p < .001$, even when controlling for Number Discrimination accuracy (Retrospective: $r(94) = .41$, $p < .001$; Prospective: $r(94) = .41$; $p < .001$). Together, both of these analyses support the conclusion that confidence sensitivity develops independently of response biases and the ANS, even when the opposite strategy children are excluded or have their data transformed.

Discussion

In Study 1, we found that children's sensitivity to confidence improves from age 3 to age 7, and that this is not fully explained by improving ANS precision or changes in response biases. We also found that both of the versions of the confidence task – the Retrospective Confidence task which asked children to evaluate their confidence *after answering* and the Prospective Confidence task which asked them to evaluate their confidence *before answering* – strongly correlated and both showed development independent of response biases or ANS precision. Finally, we found evidence in both tasks that confidence decisions are metaratio-dependent: the larger the difference in confidence between the two trials, the more likely children were to identify the more certain trial.

At the same time, however, Study 1 has two limitations. First, in order to keep children motivated in the Prospective task, we provided them with explicit feedback on their dot discrimination performance (though we gave them no feedback on the confidence portion of the task); but, in order to have children evaluate their confidence retrospectively, the Retrospective task could not give children any feedback at all. One possibility, therefore, might be that children

were trained to attend to their confidence signal throughout the course of the Prospective task and could not attend to their confidence spontaneously without feedback.

The second limitation of Study 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 & Reynvoet (2012; see also Clayton et al., 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. This leads to the possible explanation that children used distinct dimensions on the confidence and dot discrimination parts of the tasks (e.g., choosing confidence based on a cue like convex hull, but dot discrimination based on number, or cumulative area, etc.).

To test these two possibilities, in Study 2 we once again tested 3 – 7 year-old children on Prospective and Retrospective Confidence 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 Study 1, we should find that children's performance in Study 2 should be no different from chance.

Study 2

Methods

Participants. Using the correlation between age and children's confidence judgments in Study 1 (r = .40, from the Retrospective Task), we conducted a power analysis and determined that 61 participants would be required to replicate this effect with 90% power at $\alpha = .05$. 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 Study 1. One child completed only the Retrospective version, so his data was retained for analysis of the Retrospective task and removed for all other analyses.

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 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 confidence 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 confidence pairs in the same way as Study 1, with metaratios of 1.25, 1.5, 3.0, 4.0, and 5.0. All other aspects of the materials were identical to Study 1.

The procedures were the same as in Study 1 with one change: children were not given feedback about their number discrimination performance by the computer in the Prospective version. 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 Prospective and Retrospective Confidence 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 minutes on average to complete the Retrospective task and 3.1 minutes on average to complete the Prospective task.

Results

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

Number Discrimination. Replicating Study 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 a 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 Figure 2.4). Children's accuracy also strongly correlated with age, $r(59) = .66$, $p < .001$.

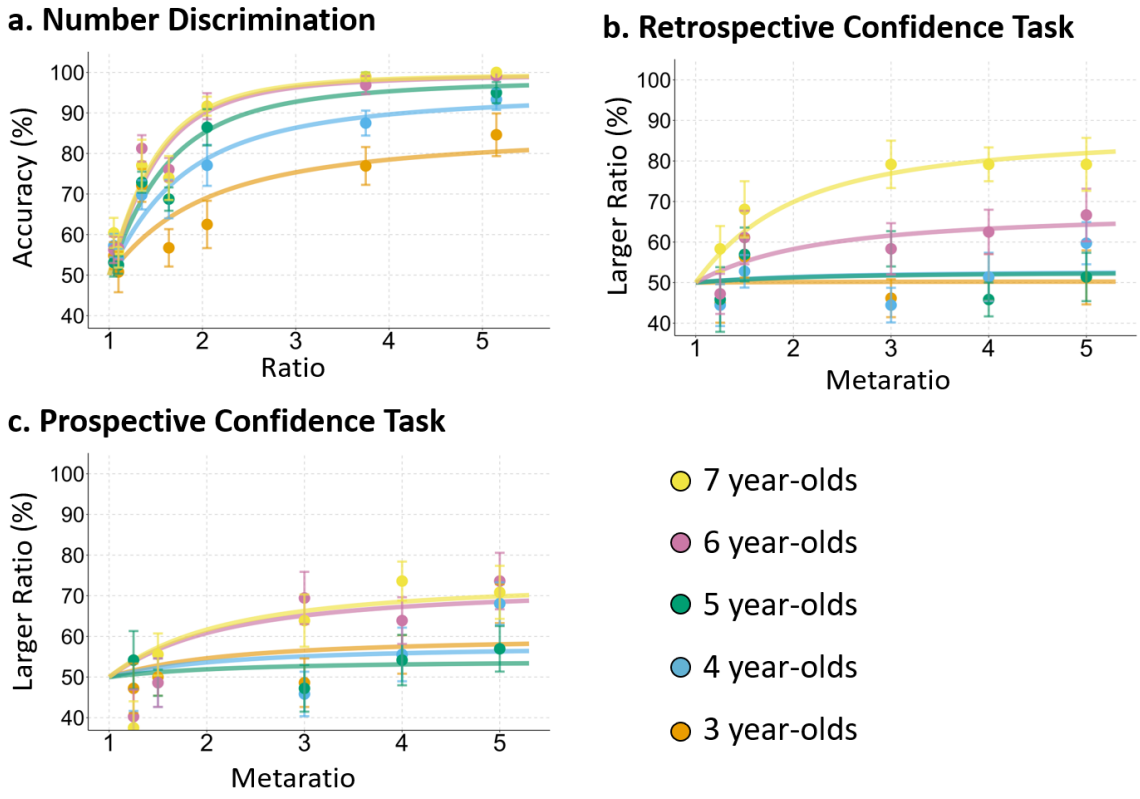
Thus, even when controlling for the five non-numeric visual features, children's performance was above chance and indicates that they relied on number to form their numerical decisions.

Retrospective Confidence Task. Children aged 6 and 7 consistently chose the trials with larger ratios above chance rates (see Table 2.2 for means and t tests). The Spearman-Brown corrected reliability estimate (with 1000 random splits) was .58, 95% CI [.41, .72]. As in Study 1, we found that children were more accurate on trials for which they indicated high confidence (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 confidence trial.

Children's confidence choice correlated with both age, $r(59) = .56$, $p < .001$, and ANS accuracy, $r(59) = .42$, $p = .001$, suggesting that their performance on the confidence 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 confidence and age did not improve the model predicting confidence 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 Figure 2.5). The correlation between age and Retrospective Confidence 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, $r(59) = .54$; $p < .001$.

Figure 2.4

Accuracy by Ratio on the Number Discrimination Trials, and by Metaratio on Confidence Trials in in Study 2.



Note. Error bars represent 1 SE, and curves are estimated using a standard psychophysical model (see Odic, 2018).

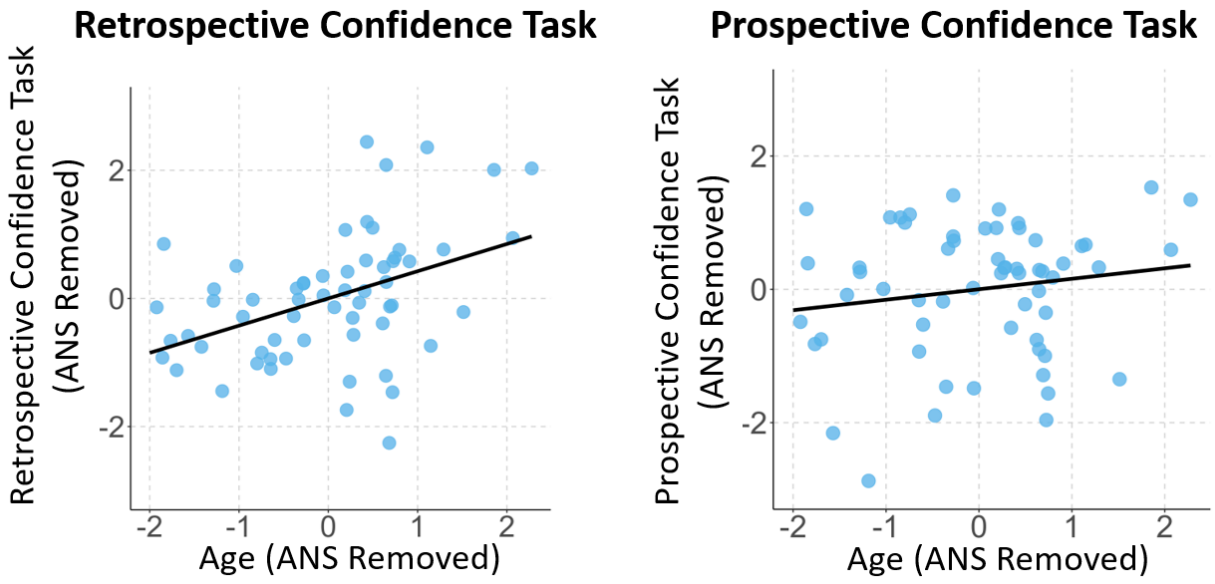
Table 2.2

Sample Sizes, Means, and Tests Against Chance for the Number Task and the Confidence Tasks in Study 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
Retrospective Confidence 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
Prospective Confidence 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

Figure 2.5

Partial Correlations Between Confidence Accuracy and Age, Controlling for Number Discrimination (ANS) Accuracy in Study 2.



Prospective Confidence Task. Replicating the Retrospective results above, children aged 6 and 7 chose to answer trials with larger numerical ratios above chance rates (see Table 2.2 for means and t tests), indicating sensitivity to their confidence. Spearman-Brown corrected reliability was .22, 95% CI [-.08, .48]. And, as in the Retrospective task, we found that a subsample of children (n = 10) consistently chose the lower confidence trial; only 4 children who went with this opposite strategy on both tasks.

Confidence choice on the Prospective 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 confidence 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 Figure 2.5). Nevertheless, we found that age and confidence 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 Study 1.

Correlations Between the Tasks. As in Study 1, children's performance on the Prospective and Retrospective 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 confidence in were nearly identical (Prospective: $M = 81\%$, $SD = 10\%$, Retrospective: $M = 79\%$, $SD = 13\%$), $t(59) = 1.03$, $p = .307$, and were correlated even when controlling for age, $r(57) = .44$, $p = .001$.

Metaratio Effects. As in Study 1, children were more likely to indicate high confidence in the larger of the two presented ratios when the metaratio between them was large, Prospective: $F(3.31, 195.52) = 13.10$, $p < .001$, $\eta_p^2 = .18$, Retrospective: $F(4, 60) = 4.61$, $p = .001$, $\eta_p^2 = .07$. There were trending interactions between age (as a covariate) and metaratio predicting children's confidence choice in both versions, Prospective: $F(3.34, 196.60) = 2.18$, $p = .083$, $\eta_p^2 = .04$, Retrospective: $F(4, 236) = 2.11$, $p = .080$, $\eta_p^2 = .04$, where older children showed metaratio effects while younger children did not (see Figure 2.4). We did not see any effect of metaratio for children's reaction times, Prospective: $F(3.09, 182.03) = 1.32$, $p = .268$, $\eta_p^2 = .02$, Retrospective: $F(3.33, 200.13) = 1.33$, $p = .263$, $\eta_p^2 = .02$.

General Discussion

Young children's ANS representations provide them not only with an approximate sense of number, but also with a sense of confidence 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 confidence 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 confidence. By testing 3-7-year-old children on two versions of the relative confidence task that directly measure sensitivity independent of criterion-setting, and by controlling for developmental improvements in children's ANS precision, we find that sensitivity in ANS confidence 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 confidence representations spontaneously, and when five non-numeric visual features are controlled for, including density and convex hull. Our findings broadly replicate claims made in the literature that children improve at monitoring their confidence with age, and

extend these claims by experimentally removing the influence of overconfidence bias and statistically removing the influence of underlying ANS noise. They also contrast to previous reports arguing that children's confidence develops primarily because of changes in response biases (i.e., calibration; Salles et al., 2016).

What, then, do our data tell us about children's sense of confidence? First, we found that ANS precision *does* correlate with confidence precision, as strongly predicted by Direct accounts and consistent with Inferential accounts. Crucially, our claim is that this Type 1 ability and the response biases we experimentally eliminated are not *sufficient* to explain confidence development by themselves. Direct accounts suggest that confidence should entirely be the product of the low-level perceptual noise, and are not easily reconciled with our data, while Inferential accounts instead suggest that the confidence signal is aggregated from a variety of sources (Koriat, 1993; Meyniel et al., 2015). These sources are sometimes proposed to include the low-level perceptual noise, 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; Meyniel et al., 2015; Pouget et al., 2016). Our data is most consistent with these aggregate models.

What, then, are the additional factors contributing to the development of confidence sensitivity beyond calibration and ANS precision? Our results are consistent with the hypothesis that the improvement in children's confidence in ANS decisions is driven not only by improvements in calibration or the ANS itself, but by improvements in the ability to reason about and represent perceptual confidence more generally. As one example, adult observers can compare states of confidence across otherwise independent perceptual boundaries, such as visual vs. auditory trials or contrast vs. orientation (De Gardelle et al., 2016; De Gardelle & Mamassian, 2014). Moreover, we find here that children's ability to reason about their confidence is ratio-dependent, providing some empirical evidence that confidence is a continuous property that itself obeys Weber's law (Weber, 1978). Together, these findings are all consistent with the possibility that confidence 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 et al., 2004).

Our study follows a tradition in the confidence monitoring literature of manipulating difficulty as a proxy for confidence because more difficult tasks intuitively should elicit less confidence. 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) as a cue for their confidence 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 Retrospective task: children were more likely to have correctly answered the trials they selected as higher in confidence than those they discarded. Second, if children were making their decisions based solely on the ratios of the two presented trials without reasoning about their subjective confidence, 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 et al., 2012; Jacob & Nieder, 2009; Matthews et al., 2016) – would account for any developmental improvements. Contrary to this, we found evidence for continuing development of confidence sensitivity when controlling for ANS precision in both the Prospective and Retrospective tasks, suggesting that the ability to reason about ratios is not the only contributing factor to children's confidence task performance.

We set out to track age-related change in ANS confidence sensitivity in the preschool and early school years, but, in both studies, we did not find strong evidence that the youngest children in our sample were sensitive to confidence. 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 confidence (e.g., Balcomb & Gerken, 2008; Call & Carpenter, 2001; Goupil & Kouider, 2016; Goupil et al., 2016; Lyons & Ghetti, 2011) and that confidence 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 confidence in children aged 5 and older, does not elicit any sense of confidence in these younger children. However, we found that 4-year-olds showed above-chance performance on the two largest metaratios in Study 1, which might suggest that these younger children *are* capable of reasoning about confidence 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, we could conceptualise confidence as being represented on a continuous and noisy domain-general scale, and perhaps young children can only discriminate large differences in confidence – larger than we presented in these tasks. At the same time, however, we failed to observe above-chance performance on the easiest ratios in Study 2 (though this could be due to the lack of feedback), leaving the factors that lead to the youngest children's success on relative confidence 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.

In sum, children can reason about their confidence in a relative task, showing development in their precision with age. Contrary to Direct accounts, age-related differences are not entirely explained by children's numerical precision, suggesting an independent maturation process for confidence monitoring. The results are, however, consistent with the interpretation of the Inferential accounts that there is information beyond what is contained in the decision that influences children's confidence judgments.

Chapter 3: The Domain-Generality of Children's Confidence

We experience confidence in a wide range of circumstances, whether that be estimating the items in our grocery basket ("am I *sure* it's less than 8?"), remembering a new colleague's name ("she said it was Sandy, *right*?"), or even evaluating an item's value ("will this cookie *really* make me happy?" [9]). Metacognitive reasoning has been studied in at least some capacity in all these decisions, though largely quarantined within their own fields (perception, memory, and neuroscience).

However, there might be reason to believe that metacognitive abilities themselves are not quarantined in such a way. For one, the findings of Chapter 2 suggest that there is development in metacognitive reasoning beyond development in the number sense. Moreover, a domain-general sense of confidence in childhood would help in part explain how children integrate and compare information across distinct and independent sources. A child trying to decide which of two groups is more socially dominant might compare their confidence in the numerical size of each group against their confidence in the physical size of each group (Pun et al., 2016).

As outlined in Chapter 1, each of the families of accounts predicts different relationships between metacognitive abilities in different domains (e.g., memory and perception), and even for metacognitive abilities in different Type 1 tasks from the same broader domain (e.g., number perception and emotion perception). Direct accounts lean towards the side of no commonalities, while Inferential accounts predict strong commonalities.

There are mixed findings about domain-generality. When examining correlations between metacognitive abilities in different Type 1 tasks, we find that confidence reasoning in perceptual tasks largely correlates within individuals (Rouault et al., 2018; Vaccaro & Fleming, 2018). Under some circumstances, such as when using measures of metacognitive sensitivity, perceptual confidence reasoning also correlates with memory confidence reasoning (Mazancieux et al., 2020). However, this effect is heterogeneous between studies (Baird et al., 2013; McCurdy et al., 2013; Rouault et al., 2018; Vaccaro & Fleming, 2018). One recent interpretation of these mixed findings is that confidence is computed by a central resource, but 'tagged' with a domain-specific interpretation to make it useful for specific systems (Rouault et al., 2018). Overall, it seems as though confidence is broader than the noise of Type 1 decisions, but perhaps not so

[9] Always.

broad as to constitute a single process in adulthood (as predicted by Mind-Reading, Cue-Based, and to some degree Bayesian accounts).

The prediction of all the theoretical accounts is that the domain-generality of metacognition should remain stable throughout development (with Direct predicting domain-specificity throughout, and Inferential predicting domain-generality throughout). And, parallel to most findings in adults, existing work has not found a correlation in children's explicit confidence judgments in math, memory, and spelling decisions below age 8 (Bellon et al., 2020; Geurten et al., 2018; and see Kelemen et al., 2000). More interestingly, metacognition even appears to be uncorrelated across independent *perceptual* dimensions in childhood. Vo and colleagues (2014), for example, measured 5-8-year-old's confidence by having them bet on getting a question right or wrong, and found that children's betting behaviour in a numerical discrimination task (e.g., which box has more dots) and their betting on an emotion discrimination task (e.g., which face looks happier) did not correlate.

The lack of correlations between metacognitive judgments on independent tasks in childhood has led some researchers to propose that metacognition may initially operate domain-specifically and gradually become domain-general (Bellon et al., 2020; Geurten et al., 2018; Lyons & Ghetti, 2010; Vo et al., 2014). This proposal is perhaps best accounted for in Inferential accounts by attributing the utility of certain cues to only a subset of metacognitive decisions. That is, children might initially learn that reaction time is generally a valid cue for memory decisions (longer thinking time means that memory is less likely to be correct), but may have to learn this separately for perceptual decisions. Over time, children would then notice that this cue is relevant to both domains and generalize the use of that cue (Lyons & Ghetti, 2010; Veenman & Spaans, 2005; Vo et al., 2014).

However, it is worth noting that the tasks reporting domain-specificity required children to explicitly rate their confidence, which, as outlined earlier, conflates children's sensitivity in their confidence representations for their response biases, such as their general tendency to overconfidently bet high. While both components contribute to confidence judgments, response biases across tasks (e.g., overconfidence in one domain and underconfidence in another, as reported by Vo et al., 2014) might influence the detection of underlying similarities in confidence sensitivity. For example, if a child is highly overconfident in their number perception, this does not necessarily mean that they will be equivalently underconfident in their emotion

perception.[10] Therefore, it remains possible that children's sensitivity to confidence representations across domains is fundamentally related, even though there was no correlation in this study. The use of the relative measure of confidence introduced in Chapter 2 could alleviate this challenge.

Here, we use the relative metacognition measure to test children's confidence acuity in three distinct perceptual dimensions: number, area, and emotion perception (Odic, 2018; Vo et al., 2014). We first confirm the domain-specificity of these dimensions in children aged 6 to 9, then investigate whether confidence sensitivity in these dimensions nonetheless correlates, signalling domain-generality, or if children's confidence representations are domain-specific before formal schooling and become domain-general as they develop.

Study 3

Methods

Participants. Eighty-one six to nine-year-old children (*M* = 7;11 [year; months], range = 6;0-10;0, 42 girls) participated in the study, meeting our a priori goal of 40 children per condition and therefore allowing for adequate psychophysical model fits (Halberda & Feigenson, 2008). Three additional children participated but were removed from the sample because they failed to complete at least 90% of the trials. Participants were tested in a quiet room at an on-campus lab or in a quiet area of their schools in Vancouver, B.C. All children spoke English and most came from middle-class families.

Materials & Procedure. Children saw custom-made stimuli on a 11.3" Apple Macbook Air Laptop using Psychtoolbox-3 (Brainard, 1997) scripts, which are available online for free use at http://odic.psych.ubc.ca/scripts/domaingeneralconfidence.zip. Children saw three types of stimuli, described in detail below and shown in Figure 3.1: blue and yellow dots (Number), a blue and yellow blobs (Area), and two emotional expressions (Emotion). Children randomly

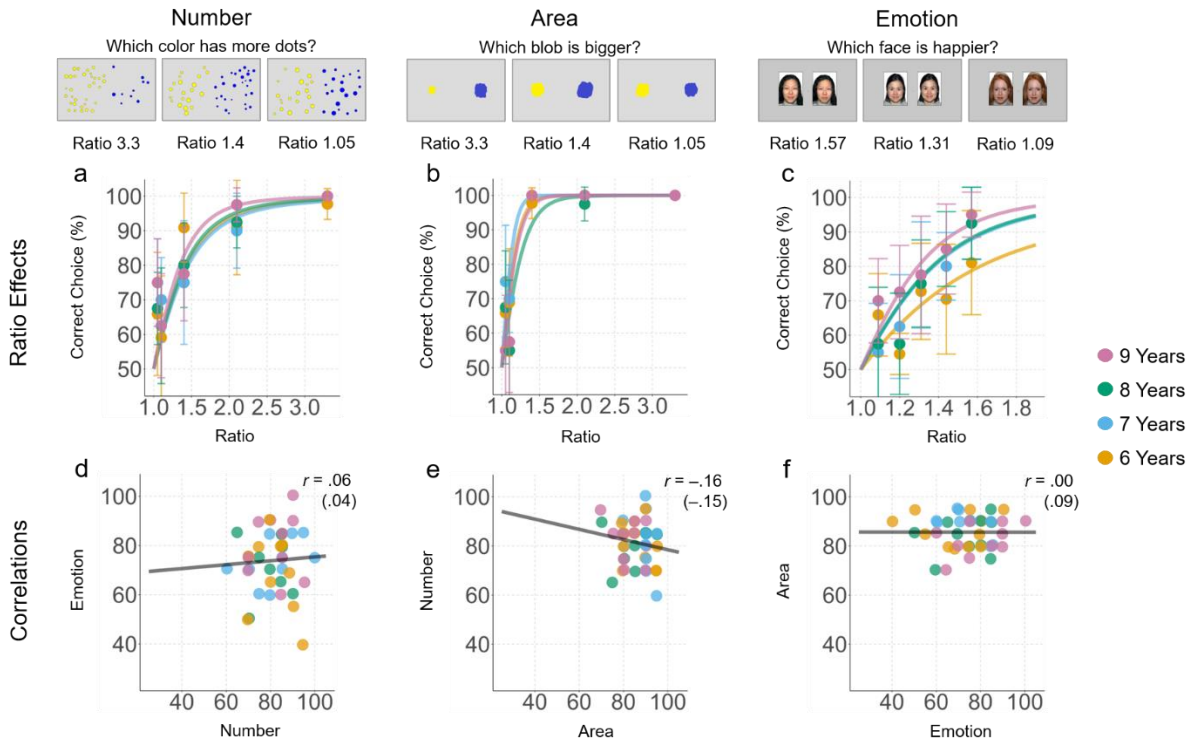
[10] Vo and colleagues (2014) tried to combat these response biases by using AUROC, one of the measures developed to separate response bias from sensitivity. However, as mentioned previously, measures like AUROC are dependent on there being sufficient trials to fully capture children's performance. In particular, the measure cannot run if a child does not bet low then get the answer correct (or bet high then get it wrong, or either of the other two combinations of confidence and accuracy). Nine of the 48 participants (19%) exhibited this behaviour by always indicating high confidence on at least one task, making it difficult to claim that the measure was able to fully account for response bias.

assigned to the Confidence condition (n = 40) were asked to reason about their relative confidence in answering two questions, while children in the Discrimination condition (n = 41) were simply asked to answer the questions. The Discrimination condition therefore allowed us to confirm the domain-specificity of the perceptual discriminations in these three dimensions, as well as control for the possibility that correlations between dimensions in the Confidence condition are due to other domain-general comparison or task comprehension abilities.

Discrimination Condition. In the Discrimination condition, children saw a Number, Area, or Emotion trial in a random, intermixed order. This both prevented order effects between the three stimuli types and made the task more interesting for children. To remove the influence of their developing motor skills and inhibitory control, children were asked to either verbalize their answer or point to one side of the screen, and the experimenter pushed a corresponding button. Children received feedback after each trial in the form of a pre-recorded female voice that would either give positive feedback (e.g., “That’s right!”) or negative feedback (“Oh, that’s not right”). Occasionally, the experimenter would give additional feedback to encourage the child to stay engaged in the task (e.g., “That’s okay, let’s do another one”). After completing twelve practice trails that familiarized children with each dimension (4 per dimension), children completed a total of 60 trials (20 per dimension).

Figure 3.1

Stimuli and Results for the Discrimination Condition of Study 3.



Note. Children were shown randomly intermixed trials from the three tasks at one of 5 ratios. As shown in panels a-c, accuracy improved as ratios increased (error bars represent 95% CI). Panels d-f show the correlations between dimensions. Correlations in brackets are controlling for age. No correlations here are significant.

Number Discrimination. This task was modelled after dozens of studies exploring children's Approximate Number System (ANS), including Studies 1 and 2. Children saw a set of yellow and blue dots, with the yellow dots on the left and the blue dots on the right (Figure 3.1) for 1000 ms – preventing them from counting – and were asked to identify "which side has more dots". We varied difficulty by manipulating the ratio of blue to yellow dots, showing children one of five ratios on each trial: 3.3 (e.g., 33 yellow dots and 10 blue dots), 2.1, 1.4, 1.1, and 1.05.

Area Discrimination. This task was modelled after studies exploring children's early ability to discriminate area (Odic, 2018). Children were shown a yellow amorphous blob on the left and a blue amorphous blob on the right (Figure 3.1) for 1000 ms and were asked to identify "which blob is bigger". We varied difficulty by manipulating the ratio of pixels in the blue and

yellow blobs, showing children one of five ratios on each trial: 3.3 (e.g., 119,130 yellow pixels and 36,100 blue pixels), 2.1, 1.4, 1.1 or 1.05.

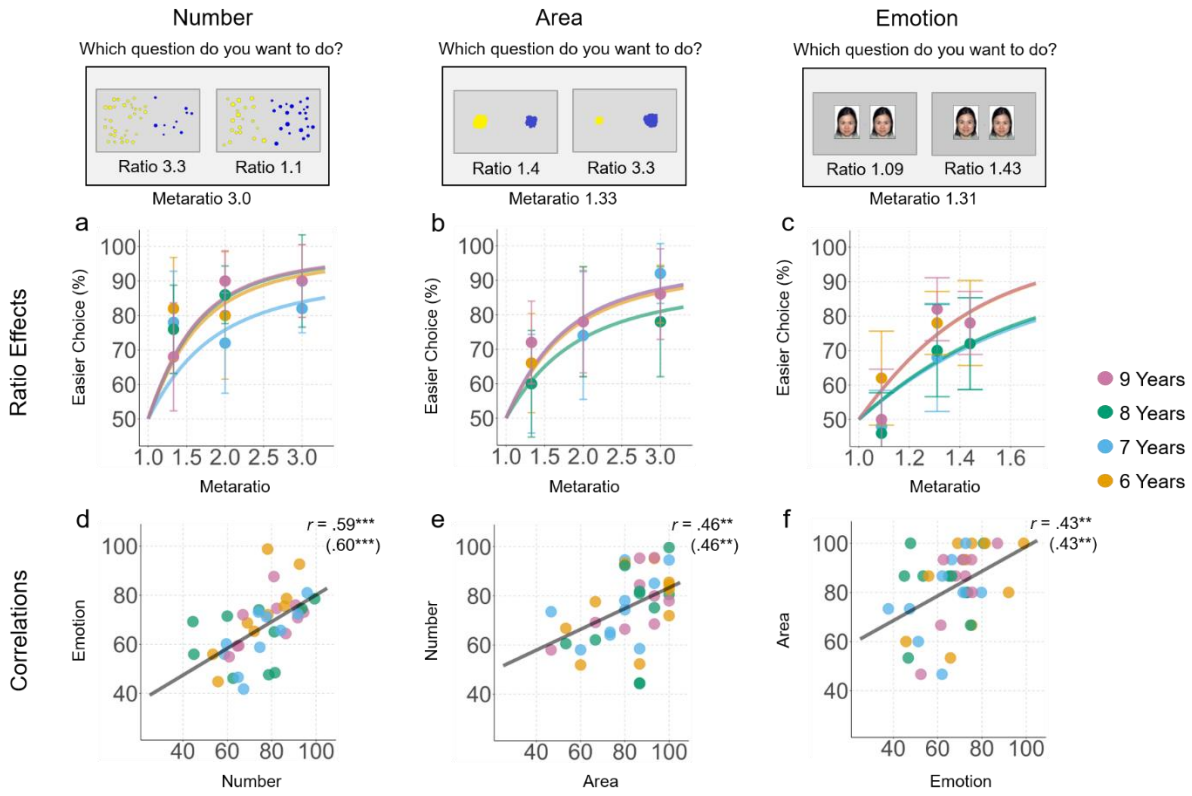
Emotion Discrimination. This task was modelled after studies exploring children's emotion discrimination (Vo et al., 2014). Children saw two female faces that differed in emotion on a spectrum from happy to angry for 1000 ms (Figure 3.1) and were asked to identify "which face is happier." To generate the stimuli, we took a 100% happy and a 100% angry face from four different female models – two Caucasian and two Asian – and blended the two faces using the FantaMorph software (version 4, Abrosoft, http://www.fantamorph.com). We blended faces in 6.67% intervals, creating 8 total blends varying from 100% happy (i.e., 0% angry), through 53.3% happy (i.e., 46.7% angry). We varied difficulty by presenting two faces whose difference was either easy to tell apart (e.g., 93.3% happy vs. 60% happy, a ratio of 1.56) and some that are very difficult (e.g., 73.3% happy vs. 66.7% happy, a ratio of 1.1). This resulted in 5 different binned ratios: 1.09, 1.2, 1.31, 1.43, and 1.57.

Confidence Condition. This task used identical stimuli to the Discrimination condition, with one simple, but major change: rather than showing children a single trial and asking them to choose the correct answer, we embedded the trials within the relative confidence task introduced in Chapter 2. Specifically, we used the Prospective version, where we presented two trials simultaneously and asked children to choose which of two trials they *wanted* to answer (Chapter 2; De Gardelle & Mamassian, 2014; Figure 3.2). As in Chapter 2, we expected that children would maximize their chances of success by choosing the more certain (i.e., easier) question. By varying the difference in difficulty between the two trials, we can identify children who can tell apart only large differences between their confidence (e.g., the difference between "very sure" and "not sure") versus children who can tell apart even small differences in their internal confidence (e.g., between "very sure" and "somewhat sure"), giving us a measure of individual differences to compare across domains.

After 12 practice discrimination-only trials, evenly distributed between each dimension, children completed 45 confidence trials (15 per dimension). On each trial, children were presented with a pair of stimuli made up of either Number, Area, or Emotion trials used in the Discrimination condition (note that children only saw pairs of stimuli from a single dimension at a time). As in the Discrimination condition, these stimuli were presented in a random, intermixed order. Children were asked to identify "which of these two questions would you like to do" and

Figure 3.2

Stimuli and Results for the Confidence Condition of Study 3.



Note. Children were shown randomly intermixed pairs of trials from one of the three tasks and asked which question they wanted to answer. Then, the screen zoomed in on their selected question and they answered it, just as in the Discrimination trials. As shown in panels a-c, accuracy improved as ratios increased (error bars represent 95% CI). Panels d-f show the correlations between dimensions. Correlations in brackets are controlling for age. ** denotes $p < .01$, *** denotes $p < .001$.

the trial would stay on the screen until children responded by verbalizing or pointing to their answer, at which point the experimenter pushed a corresponding button. To keep children motivated to choose the easier question, the selected trial would then expand to fill the screen, and children answered the question as in the Discrimination condition (e.g., judging which side has more dots in the case of Number). After choosing the answer for the selected discrimination trial, children received feedback in the form of a pre-recorded female voice. As in the Discrimination condition, the experimenter would occasionally provide additional feedback to encourage the child to stay engaged in the task (e.g., “That’s okay, let’s do another one!”), but

the child *never* received feedback on whether they had successfully selected the easier question (see Smith et al., 2008).

To vary the difficulty and estimate individual differences in the precision with which children could tell apart levels of confidence, we varied the difference in the relative difficulty between the two presented trials (i.e., the 'metaratio' – the larger numerical ratio divided by the smaller one). Children were presented with three metaratios per dimension: 3.0, 2.0, and 1.33 (for Number and Area), and 1.44, 1.31, and 1.1 (for Emotion; e.g., an easy 1.57 ratio vs. a hard 1.09 ratio yields a metaratio of 1.44).

In Studies 1 & 2, we found that children generally want to maximize their chance of success in relative confidence tasks and therefore choose the trial in which they have more confidence, producing a metaratio effect: the larger the difference between two presented ratios, the more likely children are to indicate the easier trial as the one they are more confident in. However – we also found, potentially given that we used wording that uses simple vocabulary and avoids advanced mentalistic terms (i.e., "which trial do you *want* to do", rather than "which trial are you more *confident* on"), that some of the children in our sample had significant *below chance* performance in which they consistently selected the harder of the two trials. Indeed, many of these children would subsequently tell us that they chose the harder trials in order to challenge themselves. While these children's consistent preference for the harder ratios clearly demonstrates the ability to differentiate between their confidence states (which is what we are ultimately interested in), their data also creates a bimodal accuracy distribution, leading to violations of several statistical assumptions. For this reason, all children's data were fit both by a psychophysical model assuming that children select the easier of the two trials (see Results), and an inverted version of the same model whereby we assume that children select the harder of the two trials. By using model comparison, we identified 10 children whose data clearly shows a preference for harder trials: 3 children showed this behaviour across all three dimensions, while 6 children showed this behaviour on only one of the three dimensions. For these children, we treated their inverted model as the dependent variable (e.g., a child with accuracy of 10% was modelled to have accuracy of 90%). These conclusions remain the same with these 10 children removed or modelled with the non-inverted model.

Results

We first report results from the Discrimination condition, which serves as a control condition, allowing us to detect any pre-existing correlations between perceptual representations or task understanding in number, area, and emotion perception. Subsequently, we conduct identical analyses on the Confidence condition to see whether confidence acuity correlates across the three dimensions.

Discrimination Condition. Children in the Discrimination condition performed above chance for all three dimensions (see Table 3.1 for means and tests against chance), and performed significantly better at Area compared to Number (replicating Odic, 2018), and at Number compared to Emotion, $F(2, 80) = 14.67$, $p < .001$, $\eta_p^2 = .27$. We also found that children's accuracy on the Emotion trials increased with age, $r(39) = .32$, $p = .041$, but found no significant age effects for the Area or Number trials, $r_{Num}(39) = .07$, $p = .685$, $r_{Area}(39) = -.26$, $p = .099$, most likely due to our truncated age range and small sample compared to past research in these areas (e.g., Odic, 2018). Additionally, children's accuracy varied as a function of ratio in each of the three dimensions (Figure 3.1 panels a-c): Number: $F(3.23, 129.01) = 24.66$, $p < .001$, $\eta_p^2 = .38$; Area: $F(1.95, 78.16) = 59.57$, $p < .001$, $\eta_p^2 = .60$; Emotion: $F(3.67, 146.61) = 16.45$, $p < .001$, $\eta_p^2 = .29$.

Table 3.1

Descriptive Statistics, Tests Against Chance, and Average Estimates of Fit to the Weber Model in Study 3.

Dimension	*M* [95% CI]	*t*	*p*	*d*	# Fit	*w* [95% CI]	Lapse rate [95% CI]
Discrimination Condition							
Number	81.44 [78.57, 84.30]	22.18	<.001	3.46	41	.23 [.17, .30]	.02 [.00, .03]
Area	85.58 [83.42, 87.74]	33.30	<.001	5.20	41	.10 [.10, .10]	.01 [.00, .03]
Emotion	73.86 [69.83, 77.90]	11.95	<.001	1.87	39	.30 [.16, .44]	.08 [.02, .14]
Confidence Condition							
Number	75.67 [71.05, 80.29]	11.24	<.001	1.78	37	.53 [.35, .71]	.11 [.08, .17]
Area	82.00 [77.00, 87.00]	12.95	<.001	2.05	37	.32 [.16, .48]	.13 [.07, .19]
Emotion	67.33 [63.27, 71.40]	8.63	<.001	1.36	36	.31 [.19, .43]	.21 [.13, .28]

To simplify data on each child's accuracy as a function of ratio, we fit children's accuracy data to a standard two-parameter psychophysical model widely used in the literature, yielding an estimate of each child's Weber fraction (*w*): the underlying precision of their number, area, and emotion representations independent of guessing (Halberda & Feigenson, 2008; Odic, 2018; Pica et al., 2004). Weber fractions can also be interpreted as the smallest change in a stimulus that can be reliably detected, thus smaller *w* estimates indicate better acuity. This model assumes that the underlying representations of number, area, and emotion are normally distributed tuning curves with the single parameter *w* indexing their standard deviation (i.e., precision; for review, see Halberda & Odic, 2014). On top of this standard assumption, the model fits the lapse rate, which accounts for a constant percentage of trials that participants may have been guessing (e.g., a lapse rate of 0.10 indicates that participants were randomly guessing on 5% of trials, independent of ratio). More formally, Weber fractions and lapse rates were estimated using the equation:

$$Accuracy = (1 - lapse) * \Phi\left(\frac{Ratio - 1}{w * \sqrt{1 + Ratio^2}}\right) + \frac{lapse}{2}$$

where Φ is the Gaussian cumulative distribution function. This model was fit to each participant's data for each task using R's *optim* function under the assumption of normally-distributed errors, converging on the best-fit parameters by minimizing the negative log-likelihood value. We successfully fit the data of all children except for two in the Emotion Discrimination condition (a typical cut-off of $w = 3$ was used to fit data; Halberda & Feigenson, 2008; Odic, 2018; accuracy data for these children was still included in all accuracy analyses). The estimated *w* values for the three dimensions are shown in Table 3.1, and replicate previously established values for children of this age (Halberda & Feigenson, 2008; Odic, 2018).

Finally, we found that children's performance on the three dimensions were independent of each other; how well children did on the Number trials did not correlate with how well they did on Area or Emotion, and vice-versa (Figure 3.1 panels d-f). This result held for both accuracy and *w* data, and held when we controlled for the effects of age (see Table 3.2). Thus, we can conclude, consistent with previous work (Odic, 2018; Vo et al., 2014), there is evidence for domain-specificity in children's number, area, and emotion perception.

Table 3.2

Correlations Between Tasks With and Without Controlling for Age in Study 3.

	Accuracy (*r*)		*w* (ρ)	
	Area	Emotion	Area	Emotion
Discrimination Condition				
Number	-.16 (-.15)	.06 (.04)	.02 (.02)	.00 (.00)
Area		.00 (.09)		.23 (.29^)
Confidence Condition				
Number	.46** (.46**)	.59*** (.60***)	.34^ (.34^)	.66** (.66**)
Area		.43** (.43**)		.20 (.20)

Note: Accuracy data is correlated using Pearson's r, while model fit estimates are correlated using Spearman's ρ. Correlations controlling for age are shown in brackets. ^ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Confidence Condition. Children in the Confidence condition also performed above chance for all three dimensions (see Table 3.1 for means and tests against chance), choosing the easier of the two trials on 75% of trials (95% CI[71.44, 78.78], $t(39) = 13.84$, $p < .001$, $d = 2.19$), suggesting that they reasoned about their relative confidence in the two questions. As in the case of Discrimination, we found that children were best on the Area trials and worst on the Emotion trials, $F(2, 78) = 20.50$, $p < .001$, $\eta_p^2 = .35$. And, much as in the Discrimination condition, we found no correlations between accuracy and age, all *r*'s < .09, potentially due to our restricted range.

Replicating Studies 1 and 2, we found that children were more likely to choose the easier question when the metaratio was higher, $F(2, 78) = 10.00$, $p < .001$, $\eta_p^2 = .20$. In other words, children's confidence discrimination was itself ratio-dependent. But, critically, we also found the same metaratio effect for Area, $F(1.53, 59.70) = 5.19$, $p = .014$, $\eta_p^2 = .12$, and Emotion, $F(2, 78) = 18.57$, $p < .001$, $\eta_p^2 = .32$ (see Figure 3.2 panels a-c), suggesting that it is not merely an effect of number perception, and suggesting that this confidence task can be successfully and reasonably used across a variety of stimuli types.

Because performance in the Confidence condition was metaratio-dependent, we fit children's confidence data to the same psychophysical model as the one used in the Discrimination condition, estimating each child's *confidence acuity* (the precision with which they can distinguish their internal confidence states) separately from their guessing behaviour. We successfully fit all but 8 children on all 3 tasks and all but 1 child on at least 2 tasks using the

same criteria of $w < 3$ as in the Discrimination condition, though we retained all children's accuracy data for all subsequent analyses. The fit *w* data is presented in Table 3.1.

Finally, and most importantly, we found strong correlations between Number, Area, and Emotion confidence discrimination for accuracy, and slightly weaker correlations with *w* (Figure 2, see Table 3.2). This result stands in strong contrast to the Discrimination condition and suggests an important degree of domain-generality in confidence perception that is not present when children are merely discriminating each dimension.

Principal Component Analyses. To further confirm the domain-generality of children's confidence perception, we ran two principal component analyses (PCAs), which attempt to simplify a set of variables into factors that explain the maximum possible variance (Hair et al., 2009): one for accuracy and Weber estimates for all three dimensions in the Discrimination condition, and one accuracy and Weber estimates for all three dimensions in the Confidence condition.

In the Discrimination condition, there were three components identified in the scree plot and associated Eigenvalues, clustered by dimension. To improve interpretability, the factor loadings (i.e., the correlations between variables and the extracted components) were varimax-rotated. Number, Area, and Emotion each uniquely mapped onto separate components, consistent with the interpretation that each dimension is independent (see Table 3.3 for factor loadings).

In contrast, only one component was identified in the Confidence condition, consistent with a domain-general system. Factor loadings are shown in Table 3.3 (because only one component was extracted, these could not be varimax-rotated). In sum, despite strong evidence that the underlying perceptual discriminations are domain-specific, perceptual confidence discriminations are domain-general from at least the age of 6.

Discussion

These data are the first to show evidence of domain-generality in 6-9-year-old children's sense of confidence: while children's perceptual discrimination of number, area, and emotion were dissociated, their *confidence* judgements over these same dimensions are strongly correlated and constitute a single factor, extending previous work in adults (Ais et al., 2016; De Gardelle et al., 2016; De Gardelle & Mamassian, 2014; Rahnev et al., 2015; Song et al., 2011).

We find, therefore, that children as young as age 6 share a domain-general sense of confidence with adults, suggesting that confidence is either domain-general throughout development, or else is combined before children begin formal schooling. This is not theoretically likely under a Direct account of confidence, and instead is consistent with the Inferential family of accounts.

One potential consequence of a domain-general sense of confidence is that it would allow children and adults to compare information *across* perceptual boundaries. Under many models, and consistent with our discrimination data, perceptual magnitudes are represented on distinct scales (e.g., Odic, 2018), making cross-magnitude comparison difficult. A domain-general sense of confidence could, therefore, act as a universal translator between magnitudes: given that confidence in each dimension could be represented on a scale that is shared broadly across all magnitudes, observers should be able to easily compare and decide which information is most reliable in a given context. For example, if our friend says a word that sounds like "noodle" while talking to a friend about dogs, we can use a domain-general sense of confidence to compare the auditory cues to the social cues and determine that they must have said "poodle". Similarly, an observer faced with a spontaneous discrimination task in which number is easier to discriminate than area should prioritize numerical information over other magnitudes (Cantlon et al., 2010). This prediction is strongly held by the Bayesian account, which predicts that all confidence is represented as probabilities. In support, adults can compare confidence representations between diverse perceptual domains like vision and audition, suggesting that confidence acts as a common currency within perceptual tasks (De Gardelle et al., 2016; De Gardelle & Mamassian, 2014). Similarly, confidence "leaks" from one task to another; reporting high confidence on a perceptual trial inflates the confidence rating of the subsequent trial, even if it the next trial relies on memory processing (Kantner et al., 2019; Rahnev et al., 2015).

To date, evidence of a domain-general unit of confidence has not been found in children. The Bayesian account in particular, but also other Inferential accounts, rely on the assumption that the units of confidence are domain-general throughout childhood. At their most strict, the Direct accounts also assume that the units of confidence are stable – just domain-specific rather than domain-general. Based on this assumption, Direct accounts would have to explain the presence of a domain-general unit in adulthood as the result of a learning process: over time, children take their domain-specific confidence representations and learn to 'translate' them between domains. This would mean that confidence *appears* to use a common currency in

adulthood, but only because subjects have learned by then to effectively convert one confidence currency into another (potentially from age 8-10 as suggested by Geurten et al., 2018). The Direct accounts could also argue that the correlations found in the current study could therefore be the influence of third variables (like intelligence, strategic motivation, or even just attention to the task), which at present we cannot conclusively rule out, though the lack of such correlations in the Discrimination condition suggests that these are unlikely.

Therefore, in Study 4 we provide a stricter test of whether confidence uses a domain-general unit in childhood. To alleviate the challenge of third-variable explanations, which is accentuated in the correlational analyses in the current study, we use a *within-subjects experimental* design where each child serves as their own control. We test whether children can directly compare confidence across perceptual domains, deciding whether they are more certain in one perceptual decision (e.g., number) or one from another domain (e.g., emotion; see De Gardelle & Mamassian, 2014). If confidence is represented domain-specifically in childhood, then children should be unable to compare confidence across distinct perceptual domains, or would at least be substantially worse at across- than within-domain comparisons. However, if confidence uses a domain-general common currency in childhood, then children should compare within and across perceptual domains with equivalent ease.

Study 4

Methods

Participants. Forty-eight children aged 6 and 7 years participated in the study (M = 6;11 [years; months], range = 6;0 – 7;11, 22 girls). This age group was chosen because it is younger than the reported domain-specificity at age 8 by Vo et al. (2014) and Geurten et al. (2018), but where it is known that children can make relative confidence judgments using the task parameters we were interested in (from Studies 1-3). Sample size (preregistered at http://aspredicted.org/nx5s7y) was set to be similar to related studies (e.g., Study 3; Vo et al., 2014), but anticipated to yield more power because of the entirely within-subjects experimental

design and the use of Bayesian statistical analyses[11]. Per our pre-registration plan, one child was tested but excluded from analyses for not completing the entire task. All participants were tested in the same area and manner as the other studies.

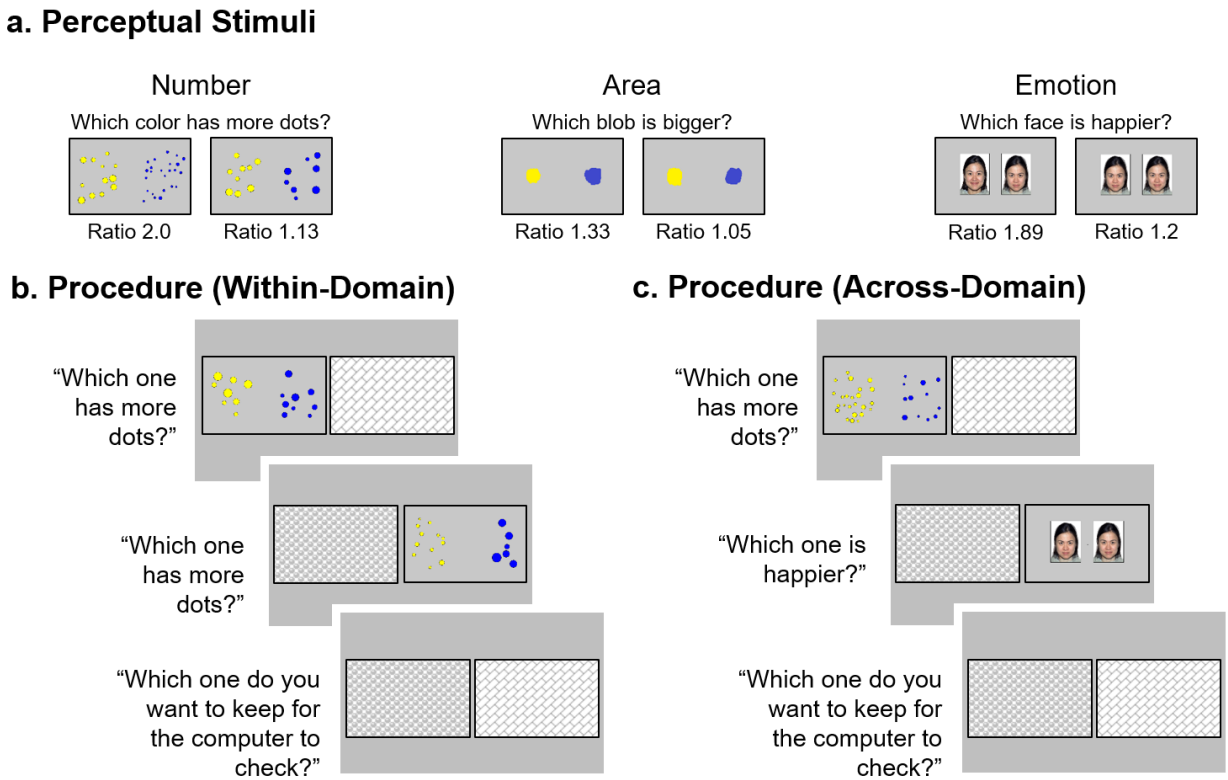
Materials & Procedure. All stimuli are available online for free use at https://osf.io/74dcv/ and were presented on an 11.3" Apple Macbook Air laptop. From these stimuli – described in detail below – we measured *perceptual accuracy/sensitivity* in number, area, and emotion decisions, and *within- and across-domain confidence accuracy/sensitivity* over these same decisions. On each trial, children saw two perceptual discrimination decisions, one easy and one hard, presented in different spatial locations on the screen. After providing an answer to both questions, children then provided a retrospective relative confidence judgment: indicating which of the two answers they were more confident in (Figure 3.3). To assess whether their confidence decisions were warranted, we examined whether the accuracy of Chosen trials (i.e., those children were more confident of) was in fact higher than Discarded trials, as would be expected if their confidence judgments were tracking meaningful information about their chances of success (see Studies 1 & 2). On two thirds of the trials, the two perceptual decisions were drawn from different domains (e.g., Number and Emotion, 'Across-Domain Condition'), allowing us to test whether children could compare their confidence across perceptual boundaries. On the remaining third of the trials, the two perceptual decisions were drawn from the same domain (e.g., Number and Number, 'Within-Domain Condition'), which served as a within-subjects control.

Following Study 3 and Vo et al. (2014), we used three perceptual domains: Number, Area, and Emotion (see Figure 3.3), which have previously been shown to be dissociable. On the Number discrimination trials, children saw groups of yellow and blue spatially-separated dots and were asked to indicate "which side has more dots." On the Area discrimination trials, children saw one yellow and one blue amorphous blob and were asked to identify "which blob is bigger." On the Emotion discrimination trials, children saw two pictures of a female face displaying a mixture of happiness and anger, and were asked to identify "which face is happier." Emotions were mixed by morphing images of happiness and anger in FantaMorph software

[11] We did not perform a power analysis *a priori* because at the time, we had not completed Studies 1 and 2, which also use the Retrospective method. Using the reported effect size from Study 1 of $d = 0.74$ in 3-7-year-old children on this task, we calculate a suggested sample size of 22 participants given $\alpha = .05$ and power = .90 to replicate the confidence effect.

Figure 3.3

Examples of Stimuli Used in Study 4.



Note: A) Sample stimuli from a high confidence and low confidence trial for the Number, Area, and Emotion tasks. B) Procedure of the confidence task for a Within-Domain comparison. C) Procedure of the confidence task for a Across-Domain comparison. Full stimuli can be found at https://osf.io/74dcv/.

(version 4, Abrosoft, http://www.fantamorph.com) for each of four female models (two Asian and two White).

To assess children's confidence, we used the retrospective version of the confidence paradigm (e.g., Studies 1 & 2; Barthelmé & Mamassian, 2009; De Gardelle et al., 2016; De Gardelle & Mamassian, 2014, see Figure 3.3), where children decide which of two trials they are more certain of answering correctly. To help children understand this task, the experimenter told children that they only needed to keep one answer for the computer to check, and so should keep the answer they were really sure they were going to get right while discarding the other answer (see Hembacher & Ghetti, 2014 for similar instructions in younger age groups). Children did not receive any feedback on the accuracy of their discrimination decisions or on their confidence

choice but were given encouraging remarks periodically to keep them motivated (e.g., “You’re going so fast!”, “Alright, let’s do another one!”).

We found in Studies 1-3 that children’s ability to compare relative confidence is best when there is a large difference in confidence between the two discrimination trials, and that confidence choices closely track accuracy. Therefore, each pair of discrimination trials had one trial that was designed to elicit high confidence (children should answer correctly 90% of the time, on average) and one trial that was designed to elicit low confidence (children should answer correctly 60% of the time, on average). Because children’s perceptual acuity in number, area, and emotion was dramatically different in Study 3, we relied on previous work mapping the developmental trajectory of perceptual acuity for each dimension to determine the key ratio that would produce 90% vs. 60% expected accuracy, thereby roughly equating the confidence strength in each domain. Therefore, for the Number trials, high-confidence trials were set at a ratio of 2.0 (e.g., 12 yellow dots vs. 6 blue dots), while low-confidence trials were set at a ratio of 1.13 (e.g., 9 yellow vs. 8 blue dots); for the Area trials, the high-confidence ratio was set at 1.33 (e.g., a 1330 px^2 blob vs. a 1000 px^2 blob) and the low at 1.05 (e.g., a 1050 px^2 blob vs. a 1000 px^2 blob); for the Emotion trials, the high-confidence ratio was set at 1.89 (e.g., a 93.33% happy face vs. a 50% happy face), and the low at 1.20 (e.g., a 70% happy face vs. a 56.67% happy face).

To evaluate whether children can compare confidence *across* domains, we presented children with two Comparison Types: Within-Domain and Across-Domain. In Within-Domain comparisons, children first saw two perceptual discrimination questions from the same domain (e.g., Number and Number), one at the high-confidence ratio and one at the low, while in Across-Domain comparisons, children saw perceptual discrimination questions from two different domains (e.g., Number and Area), one at the high-confidence ratio and one at the low.

The experiment began with twelve practice trials (four from each dimension) during which children were given instructions on how to complete the discrimination of each dimension. Afterwards, each trial began with two grey occluders on the screen. When the child was ready to begin the trial, the experimenter pushed a button to reveal the discrimination trial on the left side of the screen and asked the child the Number/Area/Emotion discrimination question. Children had unlimited time to answer, though most children did so within a second of viewing the trial. The experimenter pushed a button corresponding to the child’s decision, and

would then reveal the discrimination trial on the right side of the screen while the previous trial became occluded, again asking the Number/Area/Emotion discrimination question accordingly. After the child answered the second trial, the trial was again occluded, and the child was asked to indicate which of the two trials they just completed they would like to "keep". In total, we ran 18 Within-Domain confidence trials (6 per dimension), counterbalancing the left/right position of the easier discrimination trial, and 36 Across-Domain confidence trials (12 per pair of dimensions), making sure that each dimension had an equal number of high- and low-confidence trials. Trials appeared in a random order and were counterbalanced to ensure that each domain appeared equally on the left and right side of the screen. Note that, as a result, the experimenter did not know at the onset of any trial whether the child was going to view an Across-Domain or Within-Domain trial.

Results

Our analysis plan was preregistered, and, unless otherwise noted, all of the results reported here follow that exact plan. Because we preregistered several secondary and exploratory analyses that are not central to testing the hypothesis at hand, we only report the primary analyses of interest here.[12] None of the dependent variables interacted with age or gender (F's < 2), so we collapse across these variables in these analyses.[13] Data and annotated JASP analyses are provided at https://osf.io/74dcv/.

Discrimination Decisions. Confirming that children understood the discrimination component of each of the three tasks, children were correct on 80% (SD = 6%) of Area trials, above chance of 50%, $t(47) = 35.55$, $p < .001$, $d = 5.13$, 82% (SD = 11%) of Emotion trials, $t(47) = 20.15$, $p < .001$, $d = 2.91$, and 84% (SD = 7%) of Number trials, $t(47) = 32.60$, $p < .001$, $d = 4.71$. Reaction time data[14] revealed a slightly different pattern, with Area decisions reported fastest (M = 2780 msec, SD = 610), followed by Number (M = 2930, SD = 676), and Emotion

[12] These additional analyses can be found in the Appendix of the published article.

[13] Based on a reviewer suggestion, we found a slight side bias for children to indicate higher confidence in the right-hand answer (54%, SD = 0.11, always the second choice, see Sumner et al., 2019), $t(47) = 2.45$, $p = .018$, $d = 0.35$. Importantly, this did not differ between the Within- and Across-Domain conditions (53% Within, SD = 0.14, 54% Across, SD = 0.11), $t(47) = 0.72$, $p = .473$, $d = 0.10$, so we did not include this variable in further analyses.

[14] Note that because the children did not push the buttons themselves but instead indicated their answers verbally or by pointing, the RT measures in this case are slightly inflated from the additional time it took the experimenter to push the button or due to issues of interpretation (e.g., the child pointing at an ambiguous location). Any RTs more than 2.5 SDs from a child's own mean RT were excluded from these analyses, as we preregistered.

decisions slowest (M = 3094, SD = 804), suggesting that accuracy and RT were slightly dissociated in this sample.

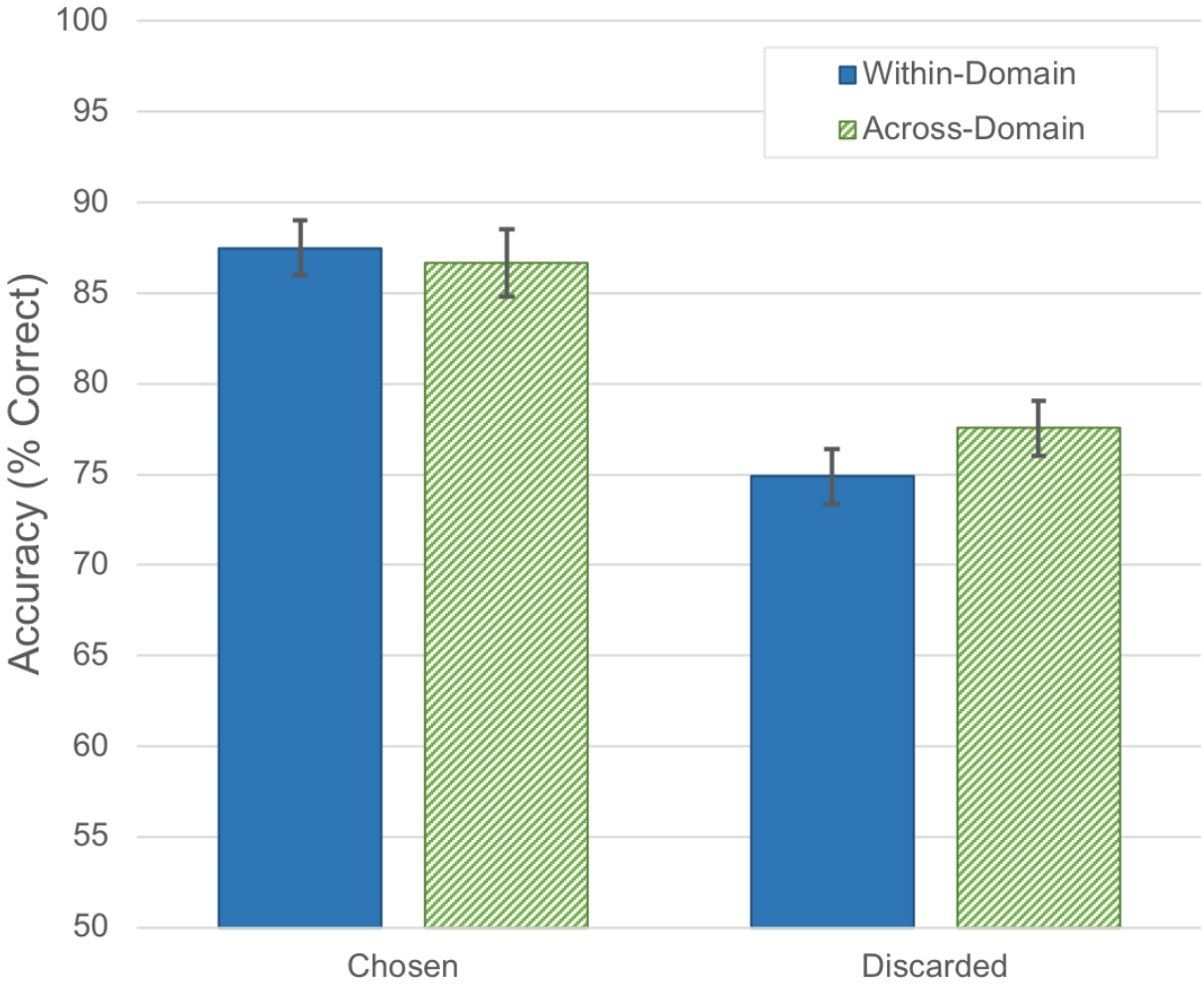
Confidence Comparison. If children made relative confidence comparisons, we should expect that their confidence choice tracks with their accuracy: "Chosen" trials should have higher accuracy than "Discarded" trials. Consistent with this, a 2 (Confidence Choice: Chosen, Discarded) x 2 (Comparison Type: Within-Domain, Across-Domain) repeated-measures ANOVA on discrimination accuracy found a significant main effect of Confidence Choice, $F(1, 47) = 30.22$, $p < .001$, $\eta_p^2 = .39$, see Figure 3.4. Thus, consistent with previous work, children's confidence choice was actually reflective of their accuracy: the Chosen trials had higher accuracy than the Discarded trials. In addition, as shown in Figure 3.4, we found no main effect of Comparison Type, $F(1, 47) = 0.75$, $p = .392$, $\eta_p^2 = .02$, nor a Confidence Choice by Comparison Type interaction, $F(1, 47) = 1.69$, $p = .201$, $\eta_p^2 = .04$; the accuracy difference in confidence choice for Within-Domain trials (M_{Chosen} = 88%, SD = 10%, $M_{Discarded}$ = 75%, SD = 13%) was not different from than in the Across-Domain trials (M_{Chosen} = 87%, SD = 10%, $M_{Discarded}$ = 78%, SD = 11%). In other words, children were *equally* sensitive at making their confidence choice across perceptual domains as within them.

Because the key finding here - that children's confidence is *not* affected by the manipulation of Comparison Type - is dependent on a null finding, we also preregistered a Bayesian repeated-measures ANOVA (using JASP with the default priors). As has been discussed at length elsewhere (Wagenmakers et al., 2018), a Bayes Factor (BF) provides the relative weight of the evidence for the null vs. the alternative hypothesis, and can therefore provide the graded strength/reliability for the null hypothesis. A BF_{10} of 1 indicates a lack of evidence for either hypothesis, while values that increase towards positive infinity indicate increasingly positive evidence for the alternative hypothesis, and values that decrease towards 0 indicate increasingly positive evidence for the null hypothesis, providing a method for testing whether a null effect is meaningful.

Conducting an identical 2 (Confidence Choice: Chosen, Discarded) x 2 (Comparison Type: Within-Domain, Across-Domain) repeated-measure Bayesian ANOVA over discrimination accuracy, we computed the model comparison for the inclusion of each of the two variables and their interaction. With a probability of .79 given the data, the best-fitting model included only Confidence Choice (BF_{10} > 1 000 000 000), providing overwhelming evidence

Figure 3.4

Children's Accuracy on Perceptual Decisions in Study 4.



Note: Results are grouped based on whether the child subsequently chose that question to keep (an indication of high confidence), or to discard (an indication of low confidence). Error bars represent 1 standard error.

that children's accuracy differed for Chosen versus Discarded trials without any influence of Comparison Type. In comparison, the model including the interaction effect of Comparison Type by Confidence Choice had a probability of .06, and was 13.68 times *less* likely than the model including only Confidence, BF_{10} = .073, providing strong evidence for the lack of an interaction. Confirming the frequentist tests reported above, we find positive evidence that children made

confidence judgments across domains as effectively as they did within domains, as predicted by a common currency account of confidence.[15]

Confidence Processing Time. A second hypothesis of a strong domain-general account is that processing confidence comparisons should not only be effective across domains, it shouldn't incur any cost in terms of processing time. That is, if children's confidence is truly represented in a domain-general format and *all* confidence judgments use these representations, then there is no processing cost to 'translate' from a domain-specific format to a domain-general format for comparisons (see De Gardelle et al., 2016). We therefore also examined whether the response time *for making the confidence decision* was equal on the Within- and Across-Domain trials.

A preregistered paired t test of reaction time found that confidence judgments for Within-Domain comparisons (M = 2175 msec, SD = 749) were faster than Across-Domain comparisons (M = 2328 msec, SD = 904), $t(47) = 2.89$, $p = .006$, $d = .42$. This was also confirmed with moderate evidence in a preregistered Bayesian analysis, $BF_{10} = 6.09$. Our data, therefore, suggest that children incurred about a 150 ms processing time cost for making confidence decisions across domains.

While a difference in reaction times might signify a processing cost of converting a domain-specific signal into a domain-general one, it could also reflect the cost of switching tasks (De Gardelle et al., 2016). If task-switching is responsible for the additional time required to complete the confidence reports, then we might expect to see a similar difference in RT for the Across-Domain discrimination decisions, as well. Specifically, the second discrimination decision made on the Across-Domain trials should be slower than the second decision in the Within-Domain trials. To examine this possibility, we conducted two non-registered exploratory analyses, finding a significant difference in RT between the second discrimination decision in the Across-Domain trials (M = 2825 msec, SD = 696) compared to the Within-Domain trials (M =

[15] A reviewer suggested that we also conduct an exploratory analysis of whether children chose one domain more than the others as an additional test of whether children were making meaningful comparisons across domains, given that children's accuracy was better on Number than on Area or Emotion. Looking only at the data from the Across-Domain trials, we found that children were more likely to select Number trials (13.33 trials of 36 Across-Domain trials, SD = 2.98) relative to Emotion (10.83, SD = 4.37) and Area (11.83, SD = 2.98), $F(1.53, 71.84) = 4.04$, $p = .032$, $\eta_p^2 = .08$ (Greenhouse-Geisser corrected). Exploratory post-hoc comparisons reveal a difference between Number and Area, $t(45) = 2.54$, $p = .043$, $d = .37$ and Number and Emotion, $t(45) = 2.50$, $p = .048$, $d = .36$ (Bonferroni-corrected). Therefore, children's confidence choices tracked the differences in accuracy between domains.

2657 msec, *SD* = 557), paired $t(47) = 2.78$, $p = .008$, $d = 0.40$. In addition, we also found that the discrimination RT cost (i.e., the difference in RT between Within-Domain second trials and Across-Domain second trials) correlated with the confidence RT cost, $r(46) = .44$, $p = .002$, strongly suggesting that task-switching costs explain the differences in confidence RTs.

General Discussion

Across two studies, we find evidence that children's perceptual confidence is domain-general from at least age 6. In Study 3, we find correlations between relative confidence choices on domains that are independent at the perceptual level. Then, consistent with the predictions of the Inferential accounts (particularly the Bayesian account), we find evidence in Study 4 that children as young as 6 years that confidence in perceptual decisions is represented using a domain-general currency: 6-7-year-old children compared their confidence equivalently well across perceptual boundaries as they did within those boundaries. This entirely within-subjects experimental method bypasses challenges faced by the correlational design of Study 3, including subjectivity to third-variable explanations, allowing us to directly test for domain-general reasoning in children. Our results thus argue against the account of developmental change from domain-specific to domain-general confidence sensitivity between 8-10 years of age. Instead, they show that confidence representations are domain-general from at least age 6, as predicted by the Inferential accounts, which all assume that confidence is domain-general throughout development.

These findings open the possibility that confidence representations might be domain-general in even younger children, as well. Preschool children, infants, and several non-human animals are known to act strategically in response to their feelings of confidence (Goupil et al., 2016; Kepecs et al., 2008; Lyons & Ghetti, 2013), just as they must do in the current paradigm. Importantly, the decisions implicated in these past findings cover a wide variety of domains: decisions about one's own memory, perceptual access to a reward, and direct perceptual comparisons like those we use here. At the same time, development in the neural structures involved in metacognitive judgments (e.g., Filevich et al., 2020) could facilitate change in the structure of confidence representations before the age of 6 tested here. Given the diversity of decisions that lead to the subjective experience of confidence in younger children, future research can test whether these decisions similarly use a common currency of confidence. This

would provide even stronger evidence of the true nature of confidence representations, and could test whether confidence reasoning is shared even among dramatically different systems like perception and memory (Rouault et al., 2018; Shea & Frith, 2019). We hope that the paradigm used here, when adapted for use with even younger children, can be a simple yet sophisticated tool for investigating questions of domain-generality in these populations.

Our findings also highlight an important distinction between sensitivity and bias in confidence reasoning (see Mazancieux et al., 2020; Winman et al., 2014). While the current studies found evidence of domain-generality in a method that experimentally isolates sensitivity, other studies found evidence of domain-specificity in studies that used calibration measures. Given evidence that bias in confidence reasoning appears domain-general in adulthood (see Mazancieux et al., 2020), *bias* may change from domain-specific to domain-general throughout development (perhaps as children gain experience in more domains or as global self-beliefs form), while *sensitivity* remains domain-general throughout. Future studies that use statistical techniques like meta-*d'* (Maniscalco & Lau, 2012), or that use a combination of measures could explore this possibility.

This work also signals the potential for domain-general transfer of confidence reasoning across perceptual tasks for children. For instance, a recent study found that training young adults' confidence sensitivity using periodic feedback about the accuracy of their confidence judgments led to improved confidence sensitivity on an unrelated task (Carpenter et al., 2019). These results hold potentially powerful implications for educational practices, as metacognitive skills are considered important for effective control of one's own learning (e.g., Lockl & Schneider, 2004). Our results suggest that similar training effects could be seen as early as primary school.

Chapter 4: Metacognition and Social Reasoning

So far, we have seen that metacognitive reasoning develops even once response biases and Type 1 abilities are controlled for (Studies 1 & 2), and that there is domain-generality among perceptual metacognition both in its processing (Study 3) and its unit (Study 4). These findings are difficult to explain under Direct accounts of confidence reasoning, as Type 1 noise should have been primarily responsible for Type 2 variability in childhood under this view. In contrast, many Inferential accounts predict that confidence is computed at one global level for all cognitive activities, using a central resource to attend to and appropriately integrate the various cues to confidence (Morales et al., 2018; Rouault et al., 2018), easily accounting for specialized development and domain-general reasoning. Nevertheless, we currently have evidence in support of domain-general metacognition only within perceptual tasks, leaving open a question of precisely how broadly confidence representations are computed.

At one end of the Inferential accounts are Bayesian theories, which critically mandate a common unit of probability among confidence decisions but remain somewhat neutral about where and how these probabilities are computed (Meyniel et al., 2015). As discussed in Chapter 3, this is critically supported by consistent evidence for a domain-general *unit* of confidence in diverse cognitive processes like visual perception, auditory perception, and memory, as evidenced by the comparability and mutual influence of confidence between independent Type 1 tasks from these domains (De Gardelle et al., 2016; Kantner et al., 2019). What remains ambiguous is whether the associated *processing*, the actual calculation of confidence, occurs domain-generally as well. There is considerable evidence for domain-general confidence processing for *perceptual* decisions (Rouault et al., 2018; Vaccaro & Fleming, 2018), even in childhood as we found in Study 3. Domain-generality between perception and other cognition, however, is much more tenuous, as outlined in Chapter 3. While some studies report dissociations in the neural correlates and behavioural signatures of confidence across domains (Baird et al., 2013; Morales et al., 2018; Rouault et al., 2018; Vo et al., 2014), some recent studies have reported strong overlap in metacognitive abilities (Mazancieux et al., 2020), and even that there is domain-general transfer of metacognitive training from perception to memory (Carpenter et al., 2019). It could therefore be possible that confidence is computed with domain-general resources that reach beyond perception.

At the other end of the Inferential view is the Mind-Reading account, which proposes that a single process is used to calculate confidence in one's own knowledge across all domains and even to reason about the knowledge of others (Carruthers, 2009; Goldman, 2006; Gopnik, 1993).[16] Effectively, children are thought to track the relationship between accurate answers and predictive behavioural cues, noticing for instance that incorrect answers tend to follow long decision times in both themselves (i.e., confidence) and in others (i.e., knowledge attributions). In some such theories, this single process primarily uses the behavioural cues associated with a decision (e.g., decision time, the experience of anxiety or pride) to infer the mental states that led to that decision both in oneself and in others (e.g., a mindreading-first account; Carruthers, 2009; Gopnik, 1993). As discussed in Chapter 1, these theories largely relied on the lack of evidence of metacognitive abilities in certain populations (e.g., children under 4, children with autism, and non-human animals) that has since been found (Elmose & Happé, 2014; Goupil & Kouider, 2019; Kepecs & Mainen, 2012). However, there are also proponents of a contrasting Mind-Reading account in which the subject uses their own metacognitive abilities to infer the mental states of others, known as the *simulation* account (Goldman, 2006; Nickerson, 1999). Under this view, interpreting others' mental states occurs through a simulation process, putting oneself in another's shoes, and attributing the experienced mental states to that other (argued to be partially achieved through the use of 'mirror neurons' that replicate others' actions; Rizzolatti & Craighero, 2004). It may therefore still be possible that a single process underlies both mindreading and metacognition – but one that is rooted in reasoning about the self.[17]

As supporting evidence for a self-first Mind-Reading account, several researchers have observed cases in which manipulating a subject's own experience affected their ability to infer the associated mental states in others. For instance, 3-month-old infants typically do not attribute goals or desires to an agent who consistently reaches towards one of two objects (Woodward, 1998), but *will* do so if given experience with grasping through the use of 'sticky mittens' (Sommerville et al., 2005). That is, only once infants have had the experience of wanting an item

[16] This literature argues that children infer mental states and knowledge, but for our purposes it is also sufficient to consider a simpler process of predicting another's accuracy. For consistency with the established literature, I refer to this as interpreting mental states/knowledge.

[17] Note that this account assumes that subjects already have metacognitive abilities to some degree, being capable of monitoring their own internal states, and therefore is not itself an explanation for how confidence is computed. Rather, this is one account of how broad metacognitive abilities could be useful not only in reasoning about the self but about others, too.

and being able to grasp it (something difficult for 3-month-olds because of slow motor development) can they then reason about the presence of similar mental states in others. Within the realm of confidence judgments specifically, when adults and children make judgments of learning (a type of confidence judgment based on one's expectations about how successful future performance on an item will be), they are more accurate in predicting others' learning if they have first evaluated their own performance (Koriat & Ackerman, 2010b; Paulus et al., 2014). In fact, reasoning about others can go awry if the metacognitive process itself is misled: adults estimated that a larger proportion of their peers would know the answers to trivia items that they were taught but had forgotten relative to items they were never taught (Birch et al., 2017). In both cases, the participant was ignorant of the answers but the familiarity of the questions in the first case (or a feeling of *fluency*) was misattributed as an indicator of high confidence (and see Haddock et al., in preparation for similar findings in childhood). These findings therefore suggest that metacognitive processes could be calculated more broadly than within perception alone or within memory alone and even encompass parts of social reasoning.

At a basic level, one would expect that if confidence is calculated broadly enough to facilitate reasoning about the strength of one's own knowledge and the strength of another's knowledge, reasoning about own and other's knowledge strengths should correlate. In support of this prediction, one recent study reported that 18-month-old infants who persisted in searching for a toy after making an incorrect choice about its location (a sign of poor metacognitive monitoring) were also more likely to learn a new object label from an unreliable teacher (Kuzyk et al., 2020). By using a measure of persistence, this study avoided the confound of developing mental state language that was present in past work (e.g., Gopnik & Astington, 1988). However, as both tasks relied on absolute judgments (choosing to keep searching or not and the choice to believe the label or not) the study does not account for confounding response biases. That is, when faced with uncertainty (like a new word or the location of a toy), some children might be naturally inclined to seek out as much information as possible from the world both by searching longer or by trusting adults, leading to correlations that are driven entirely by children's information-seeking biases and not their metacognitive abilities, per se. To eliminate this possibility, we can once again use the relative confidence task from Studies 1-4 to capture individual differences in reasoning about one's own knowledge strength independent from both developing mental state language and response biases.

To capture individual differences in children's evaluation of the strength of others' knowledge, we turned to the literature on children's selectivity in learning, in which children are well-documented as attending to past accuracy to inform future learning (essentially the equivalent relative/2AFC task for social reasoning in children). Traditionally in these studies, children are presented with two potential teachers who provide conflicting information and must pick who to believe (Birch et al., 2008; Koenig & Harris, 2005b; Mills, 2013; Tong et al., 2020). One such study found that children as young as 4 years old were sensitive to fine-grained differences in accuracy, much the same way we detected fine-grained sensitivity to own accuracy in Studies 1-4. When two informants both incorrectly estimated the number of dots on a card, children selectively preferred to learn from the informant whose guess was *closer* to the true number (Einav & Robinson, 2010). Because this study used fine-grained differences in accuracy rather than all-or-none evaluations, we felt that this task would naturally capture individual differences in children's ability to reason about others' accuracy.

If metacognitive reasoning about confidence is computed at a domain-general level that includes reasoning about the strength of others' knowledge (predicted by the Mind-Reading accounts), then we expect to see strong correlations between the relative metacognition task and the relative selective social learning task, especially given that the two reduce the influence of response biases compared to other metacognitive and social reasoning tasks. However, if computations of confidence are not done at this level of domain-generality, then we expect to see no correlations between these two tasks.

Study 5

Methods

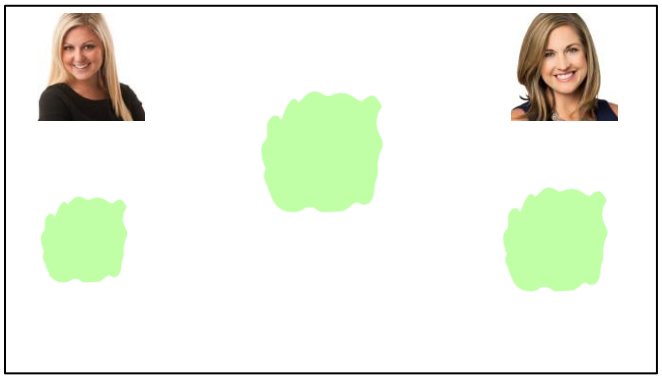
Participants. A total of 81 children participated in the study (M = 5;11 years;months, range = 4;0 - 7;10, 50 girls), meeting the planned sample size of 80 children (20 per age group, set arbitrarily a priori). Children were tested individually in a quiet area of their schools and daycares. All children spoke enough English to carry on a short conversation, and were predominantly White or South-East Asian and middle-class, as is representative of the Vancouver area.

Materials and Procedures. Children completed two tasks in a fixed order: a selective social learning task and a confidence discrimination task. Both tasks were presented on a laptop.

Selective Social Learning Task. We created a modified version of the selective social learning task used by Einav and Robinson (2010) in which two informants provide the child with incorrect answers, but one answer is relatively more accurate than the other. Children saw photos of a pair of "contestants in a drawing contest" and were asked to help the experimenter "choose

Figure 4.1

Selective Social Learning Stimuli used in Studies 5-7.



Note. In the center is the target object, flanked by each contestant's reproduction (the right contestant is the 'closer' contestant in this example). Children saw three trials like this before making their Winner, Ask, and Endorse judgments about this pair of contestants.

the winner". At the beginning of the study, the experimenter told children that the contestants had to copy a shape perfectly, and that it was particularly important for the shape to be the same size (children were asked to repeat this rule to ensure understanding). We felt that using differences in size rather than numerosity, as Einav & Robinson did, would make the task more accessible to children who were still learning number words, thereby eliminating the need for a number line to cue children to the relative placement of the informant answers.

To obtain coarse individual differences, we had children complete 4 trials, each with three judgments of informant accuracy. In each trial, children were introduced to a new pair of White female contestants (shown in photographs rather than videos, see Kominsky et al., 2016 for a similar method) and saw three examples of their drawings which critically differed only in size. Each example started with the shape in its target size in the center of the screen, followed by the "copies" made by each contestant underneath their respective photos (see Figure 4.1).

Across the three examples, one contestant consistently produced shapes that were closer in size to the target than the other contestant's shape. Shapes varied in whether they were too large or too small between examples to avoid children learning a rule that the smallest/largest shape was always the winner. The closer-sized shape was either larger or smaller than the target by a ratio of 1.2 (e.g., 120% or 83% of the original size), and the further-sized shape was a ratio of 2.0 (e.g., 200% or 50%) over or under in the same direction. Across the 4 trials, the left/right positioning of the closer-sized shape was counterbalanced, and the identities of the 'winning' contestants were counterbalanced between participants.

Following the three examples in each trial, children answered three test questions based on classic selective social learning measures. First, the experimenter asked children to choose the "Winner" of the contest (i.e., who drew their shapes closer in size to the targets), providing a direct assessment of whether children detected the difference in accuracy. Children were then told about a drawing contest in their class later, and they could ask one of the two contestants to come help them, allowing us to see if their assessments of accuracy carry over to their judgments of worthy teachers (an "Ask" judgment). Finally, the experimenter pretended as though she was showing another example drawn by the contestants, but the target shape didn't show up because of a "computer glitch". Instead, children saw shapes of different sizes drawn by the two contestants and were asked to indicate which shape was probably more like the target [that didn't show up]. Thus, much like "Endorse" trials in other studies (e.g., Koenig & Harris, 2005), we expected children to rely on their previous judgments of competency to make their selection, given the absence of an objective answer.

Confidence Task. To assess individual differences in children's sensitivity to confidence, we administered the area version of the Confidence Discrimination task from Study 3. In the area task, children select which of two shapes is bigger (see Odic, 2018). Critically, we manipulated the degree of confidence participants should feel in the area questions by manipulating the ratio of pixels in the blue and yellow blobs. Larger ratios, like ratio 3.3 (e.g., 119,130 yellow pixels and 36,100 blue pixels), consistently lead to higher confidence than smaller ratios like 1.05 (see Studies 1-4). We used 5 ratios in total for this task: 3.3, 2.1, 1.4, 1.1, and 1.05.

To assess whether children were sensitive to these differences in confidence, we showed children screenshots of images from the area task in pairs on the screen and selected which of the two screenshots they "wanted to answer", after which they answered only the selected question.

Screenshots were then paired to make three "metaratios": differences in difficulty between the two screenshots. For example, one trial with a ratio of 3.3 on the left and a ratio of 1.1 on the right yields a metaratio of 3.0 (3.3 / 1.1). Children were presented with three metaratios: 3.0, 2.0, and 1.33. By varying the difference in difficulty, we can identify children who can tell apart only large differences between their confidence (e.g., the difference between "very sure" and "not sure") versus children who can tell apart even small differences in their confidence (e.g., between "very sure" and "somewhat sure"), yielding a measure of individual differences.

As we saw in past studies, approximately 90% of children strategically choose the easier of the two images, relying on a subjective sense of their own confidence in being able to correctly answer the question. The remaining 10% of children strategically choose the *harder* of the two images, often citing a desire to challenge themselves. While the response produced by these children is qualitatively different, the underlying ability to detect differences in confidence that we are interested in (i.e., to identify which trial is 'easy' and which is 'hard') remains identical to children who chose the easier option. In fact, given that our analyses rely on correlations, these children could artificially induce correlations where none otherwise exist. However, these children can objectively be identified using a psychophysical model presented in Study 3 that expects children to be more likely to select the easier question as the difference in difficulty increases. If children become *less* likely to select the easier question (i.e., consistently pick the harder question), their data will only fit an inverted model, allowing us to identify these children and invert their data.

The Confidence Discrimination task consisted of 15 trials (three at each metaratio), with 4 warm-up trials of the area task alone. Children received pre-recorded feedback from the computer ("Yeah, that's right!", "Oh, that's not right.") when they answered the area question, but received no feedback about whether they chose the easier question.

Results

Selective Social Learning Task. First, we examined whether children's performance on the Selective Social Learning task replicated typical patterns in each of the three response types: Winner, Ask, and Endorse. At ages 5, 6, and 7, children reliably chose the contestant who made lesser errors as the Winner, though children at age 4 did not (see Table 4.1 for means and tests against chance of 50%). Performance correlated with age, $r(79) = .56$, $p < .001$, indicating that

older children were better at detecting and attributing the differences in sizes to the contestants. Similarly, children aged 5 and 7 reliably Asked the contestant with lesser errors for help, while children aged 4 and 6 did not, see Table 4.1. There was still a significant correlation with age, $r(79) = .38$, $p = .001$, suggesting that older children were more likely to use their judgments of who committed lesser errors to inform their help-seeking decisions. However, we found a very different pattern of results in children's Endorsement of one contestant over the other when lacking an objective reference: all age groups chose at chance rates, with no difference between age groups, $r(79) = .16$, $p = .153$. This was unexpected given that a recent meta-analysis conducted on similar tasks found strong evidence that children endorse informants they believe to be more accurate (Tong et al., 2020).

We then examined whether the three responses were related to one another, that is, whether children would consistently choose the same contestant across the three trials (e.g., Einav & Robinson, 2010; Koenig & Harris, 2005a). Consistent with previously published findings, children's Ask choices were correlated with their choice of Winner when controlling for age, $r(78) = .58$, $p < .001$. Children's Endorse choices were not correlated with their choice of Winner, $r(78) = .11$, $p = .339$, and only marginally with their Ask choices, $r(78) = .21$, $p = .067$ when controlling for age. Overall, then, we found that children deemed the closer contestant the Winner and were more likely to Ask her for help, but did not necessarily Endorse her on a later trial.

Confidence Task. In this task, we expected children to select the easier of the two screenshots, if they could tell them apart using their subjective confidence. Accordingly, we found that children at all age groups selected the easier trial more than 50% of the time, see Table 4.1 for means and tests against chance.[18]

Because this is the first time the area version of this task has been used with children under 6, we also examined whether children's performance depended on the metaratio – the difference in difficulties between the two presented trials. In a repeated-measures ANOVA on the three metaratios (1.33, 2.0, and 3.0) with Age Group as a between-subjects factor, we did not

[18] Note that for these analyses, children who responded consistently with the harder option (11 children, detected by the psychophysical model) have been inverted to match the response pattern of the rest of the sample. However, with their original data, this test against chance remains significantly above chance in 4, 5, and 6-year-olds.

Table 4.1

Means and Tests Against Chance for the Selective Social Learning Responses, Confidence Task, and Area Task in Studies 5-7

	Study 5					Study 6					Study 7					Mega-Analysis				
	Mean (%)	SD	*t*	*p*	*d*	Mean (%)	SD	*t*	*p*	*d*	Mean (%)	SD	*t*	*p*	*d*	Mean (%)	SD	*t*	*p*	*d*
Winner																				
4	40.00	26.16	-1.71	.104	0.38	51.19	23.02	0.24	.815	0.05						45.73	24.94	-1.10	.280	0.17
5	64.29	26.89	2.43	.024	0.53	66.25	24.70	2.94	.008	0.66	62.22	29.01	2.83	.007	0.42	63.66	27.29	4.64	< .001	0.50
6	86.25	28.65	5.66	< .001	1.27	63.75	30.86	1.99	.061	0.45	79.55	26.56	7.38	< .001	1.11	77.38	28.94	8.67	< .001	0.95
7	86.25	23.61	6.87	< .001	1.54	78.75	20.32	6.33	< .001	1.41	94.19	14.26	20.31	< .001	3.10	88.55	19.25	18.25	< .001	2.00
Ask																				
4	47.50	22.80	-0.49	.629	0.11	54.76	24.52	0.89	.384	0.19						51.22	23.68	0.33	.743	0.05
5	67.86	23.90	3.42	.003	0.75	58.75	27.24	1.44	.167	0.32	54.44	29.33	1.02	.315	0.15	58.72	27.86	2.90	.005	0.31
6	63.75	33.91	1.81	.086	0.41	60.00	32.85	1.36	.189	0.30	77.27	25.75	7.02	< .001	1.06	69.94	30.25	6.04	< .001	0.66
7	80.00	28.79	4.66	< .001	1.04	70.00	25.13	3.56	.002	0.80	88.95	18.34	13.93	< .001	2.12	82.23	23.92	12.28	< .001	1.35
Endorse																				
4	41.25	24.70	-1.58	.130	0.35	47.62	24.88	-0.44	.666	0.10						44.51	24.69	-1.42	.162	0.22
5	46.43	31.90	-0.51	.614	0.11	47.50	25.52	-0.44	.666	0.10	50.00	26.11	0.00	1.00	0.00	48.55	27.21	-0.05	.622	0.05
6	52.50	35.26	0.32	.755	0.07	55.00	27.63	0.81	.428	0.18	64.77	31.12	3.15	.003	0.47	59.52	31.50	2.77	.007	0.30
7	48.75	32.92	-0.17	.867	0.04	51.25	28.65	0.20	.847	0.04	66.86	28.72	3.85	< .001	0.59	58.73	30.60	2.60	.011	0.29
Confidence																				
4	65.00	12.40	5.41	< .001	1.21	59.76	8.29	5.40	< .001	1.18										
5	68.57	15.94	5.34	< .001	1.17	63.50	12.99	4.65	< .001	1.04	68.89	14.00	9.05	< .001	1.35					
6	66.67	14.18	5.26	< .001	1.18	78.50	13.19	9.66	< .001	2.16	71.97	13.83	10.53	< .001	1.59					
7	69.33	14.73	5.87	< .001	1.31	77.75	11.29	10.99	< .001	2.46	74.11	12.87	12.29	< .001	1.87					
Area																				
4						75.71	13.65	8.63	< .001	1.88										
5						82.62	5.29	27.60	< .001	6.17	83.56	9.51	23.66	< .001	3.53					
6						84.88	7.88	19.78	< .001	4.42	86.7	8.69	28.02	< .001	4.22					
7						84.25	3.98	38.47	< .001	8.60	86.05	8.13	29.06	< .001	4.43					

find evidence of any effect of metaratio, $F(2, 154) = 2.49$, $p = .087$, $\eta_p^2 = .03$, age group, $F(3, 77) = 0.37$, $p = .775$, $\eta_p^2 = .01$, or their interaction, $F(6, 154) = 0.66$, $p = .681$, $\eta_p^2 = .02$. We therefore do not find strong evidence that the difference in difficulties affected children's choices.

Correlations of Individual Differences. Given that both tasks worked as expected (except the Endorse trials in the Selective Social Learning task), we next looked for correlations of individual differences between the two tasks. As shown in Table 4.2, the Confidence task correlated with both the Winner and Ask choices, but not with Endorse choices. These effects held when controlling for age.

Table 4.2.

Correlations Between Selective Social Learning Measures and Confidence Discrimination in Studies 5-7.

Study	Winner	Ask	Endorse
Study 5	.30**	.38**	.00
Controlling Age	.29*	.36**	-.02
Study 6	.31**	.19	.20
Controlling Age	.14	.08	.12
Controlling Area	.28*	.16	.17
Controlling Age and Area	.14	.07	.14
Study 7	.06	.07	.06
Controlling Age	-.01	-.01	.03
Controlling Area	.06	.07	.06
Controlling Age and Area	-.02	-.01	.03
All Studies	.21***	.20**	.08
Controlling Age	.09	.1	.04
Controlling Area	.21**	.19**	.08
Controlling Age and Area	.09	.1	.04
Controlling Age, Area, and DVAP	.08	.09	.03

*Note: * denotes p < .05, ** denotes p < .01, *** denotes p < .001*

Discussion

We found a small but significant correlation between two existing measures of sensitivity in selective social learning and in confidence reasoning, consistent with the prediction that selectivity in social learning makes use of confidence reasoning. We also generally replicated the key findings of Einav & Robinson (2010) using a modified procedure, albeit not for 4-year-olds or in Endorse trials, and we found for the first time that children under age 5 could reason about their relative states of confidence.

However, both the Selective Social Learning task and the Confidence task relied on size comparisons, meaning that at least some part of this correlation could be accounted for by children's ability to reason about area. Therefore, in Study 6, we sought to replicate these findings while also controlling for individual differences in area reasoning by using the Retrospective Confidence reasoning task that allowed us to simultaneously collect data on children's area sensitivity.

Study 6

Methods

Participants. Eighty-one children participated in the study (*M* = 6;0, range = 4;0 - 8;0, 39 girls), in the same manner and geographical location as Study 5. Two additional children were tested but not included in the sample because they did not complete the study. None of the children had participated in Study 5. In addition to the two tasks below, we asked parents to complete a short vocabulary assessment online in the two weeks following participation in the study (the Developmental Vocabulary Assessment for Parents or DVAP; Libertus et al., 2015), but we had low rates of completion (26 of 81 participants). We report these data in the mega-analysis presented following Study 7.

Materials and Procedures.

Selective Social Learning Task. We used the same 'drawing contest' task from Study 5, but with two small changes. First, to increase the variability of individual differences and potentially make the task possible for 4-year-olds, we modified the degree of error in two of the four trials. As in Study 5, two trials featured errors at a ratio of 1.2 (e.g., 120% or 83% of the original size) against 2.0 (200% or 50%), while the other two trials featured a ratio of 1.2 against

3.0 (300% or 33%). Second, we made the two shapes in the Endorse trials exactly the same size to reduce the likelihood that children would rely on alternative heuristics like 'choose the largest shape'.

Confidence Task. We used the Retrospective version of the confidence task from Study 5, in which children make relative judgments about which of two questions makes them feel more certain, but this time the judgment is made *after* they have answered both questions (see Studies 1, 2, & 4). Like before, children saw four warm-up trials of the 'blobs game' before being introduced to the confidence portion of the task. The experimenter told children that they would need to get a lot of questions correct in order to win the game, but that the child could choose between pairs of questions and keep the answer they were "more sure" they got right. On each trial, children answered one question on the left side of the screen, then one question on the right side of the screen, and then made a choice about which one they were most sure they got correct. Questions were never visible on the screen at the same time. Area discrimination questions ranged from a difficult ratio of 1.03 to and easy ratio of 3.3 and were paired to form 20 confidence trials at metaratios of 1.1, 1.33, 2.0, and 3.0. Because all children had to answer the same 40 Area questions, we used their accuracy on these questions as a measure of area precision. Children did not receive feedback in any part of the task.

Results

Selective Social Learning Task. Replicating Study 5, children aged 5 and older chose the informant with lesser errors as the winner more often than chance of 50% (see Table 4.3 for means and tests against chance), and choice correlated with age, $r(79) = .35$, $p = .001$. However, only 7-year-old children chose to Ask this contestant for help significantly more than chance (see Table 4.3), also correlated with age, $r(79) = .22$, $p = .048$. Once again, no age group Endorsed the closer contestant above chance rates (see Table 4.3), with no correlation with age, $r(79) = .07$, $p = .528$. As in Study 5, children's choice of Winner was related to who they Asked for help, $r(78) = .43$, $p < .001$, and Ask judgments correlated with Endorse judgments, $r(78) = .28$, $p = .013$, but Endorse responses did not significantly relate to choice of Winner, $r(78) = .19$, $p = .085$ (all correlations controlling for age). Therefore, despite attempts to make the task easier for younger children and to make the Endorse question clear, we found that many children in the

sample did not use their judgment of who made lesser errors to inform their Ask and Endorse choices.

Area Task. Children at all ages accurately chose the larger shape well above chance levels (see Table 4.3), and accuracy increased with age, $r(79) = .33$, $p < .001$. Performance on the area task was not significantly correlated with children's answers on Winner, $r(79) = .13$, $p = .260$, Ask, $r(79) = .12$, $p = .278$, or Endorse questions, $r(79) = -.06$, $p = .595$.

Confidence Task. Replicating the results of Study 5, children at all ages chose the easier trial as their most certain more often than expected by chance (see Table 4.3), with a significant correlation with area discrimination, $r(79) = .28$, $p = .011$. In an ANOVA on the 4 metaratios (1.1, 1.33, 2.0, and 3.0) with age group as a between-subjects factor, we found that children were more likely to choose the easier question as they got older, $F(3, 77) = 14.16$, $p < .001$, $\eta_p^2 = 0.36$, and were more likely to choose the easier question as the difference in difficulties increased, $F(3, 231) = 16.65$, $p < .001$, $\eta_p^2 = .18$. There was no interaction between the two, $F(9, 231) = 0.45$, $p = .907$, $\eta_p^2 = .02$.

Replication of Study 1 & 2. Because we had a measure of both area accuracy and confidence choices, we also looked for two effects shown in number confidence decisions in Studies 1 & 2: higher area accuracy on trials children chose to keep (e.g., high confidence) compared to discarded trials (low confidence), and age-related differences in confidence performance independent of area performance. Replicating Studies 1 & 2, we found that accuracy on high confidence trials ($M = 86.23$, $SD = 12.59$) was higher than accuracy on low confidence trials ($M = 77.35$, $SD = 10.99$), $t(80) = 5.44$, $p < .001$, $d = 0.75$.[19] We further found that adding area performance to a hierarchical regression including age did not explain any additional variability in children's choice of the easier trial, $R^2 = .34$, $F(2, 78) = 19.96$, $p < .001$, $R^2_{Change} = .01$, $F(1, 78) = 1.10$, $p = .297$, $\beta_{Age} = .54$, $t(78) = 5.53$, $p < .001$, $\beta_{Area} = .10$, $t(78) = 1.05$, $p = .300$. Together, these findings replicate all the key findings from the number confidence task, but with area decisions as well.

Correlations of Individual Differences. Alone, the confidence measure correlated with children's choice of Winner, but not their Ask or Endorse answers (see Table 3.2). However, when controlling for age and performance on the area task, this effect became non-significant.

[19] Note that for this analysis, we did not invert any data based on the psychophysical model.

Discussion

While we replicated some results of Study 5 in that children were selective in who to trust, were sensitive to differences in confidence, and these two correlated, we saw two major differences. First, only 7-year-olds, the oldest children in our sample, showed above-chance selectivity to ask the closer informant for help. This contrasts with the selectivity of even 5-year-olds choosing the closer informant as the winner, with Study 5 where we found selectivity in 5-year-olds' Ask choices, and with Einav & Robinson's second study (2010) where they found selectivity in 4-year-olds' Ask choices. We will note that though the effects did not reach significance, they were in the direction we predicted, potentially meaning that we lacked sufficient power to detect these effects. Second, we did not find that performance on the confidence task correlated with any of our selective social learning measures when controlling for age and area. We found that while children's performance on the Retrospective Confidence task was consistent with previously reported findings, there was a large correlation with age that was not found in Study 5. It may therefore be possible that by controlling for age in this task, there was insufficient variability left to detect any correlation. As potential evidence for this interpretation, the correlation between children's choice of Winner and their performance on the confidence task when controlling only for area was about the same magnitude as in Study 5 and was trending significance (see Table 4.2).

Given these possible challenges for interpreting this data, we conducted one additional study using a larger sample of children (aged 5 and older only, given the lack of meaningful performance on the selective social learning task by 4-year-olds in both studies) using the Prospective Confidence task used in Study 5 to avoid the issue of diluted variability when controlling for age, and including a short assessment of area discrimination abilities.

Study 7

Methods

Participants. Using the observed correlations in Studies 5 & 6 between children's choice of Winner and performance on the Confidence task, we calculated a sample size of 129 children would allow us to detect an effect with .90 power at $\alpha = .05$. This sample size is also sufficient to

detect an effect as small as $d = .28$ when comparing children's selective social learning answers against chance. Rounding this sample up to counterbalance our stimuli, we tested 132 children (M = 6;5, range = 5;0 - 7;11, 82 girls) in the same manner as Studies 5 & 6. Three additional children were excluded for not completing the study in full, and none of the children in the sample had participated in Studies 5 or 6. As in Study 6, we asked parents to fill out the DVAP online within two weeks of participation. Forty-seven parents completed the assessment, and these data are reported in the mega-analyses.

Materials and Procedures.

Selective Social Learning Task. We used the same stimuli as in Study 6. The only change was in the wording of the "Endorse" question, now asking children "Which girl's shape would you guess looks the way it is supposed to look?"

Area Task. Immediately after the Selective Social Learning task, children completed a 20-trial area discrimination task that served as a control for the similarities between the two key tasks. The task used the same type of stimuli as the confidence discrimination task (e.g., children chose the larger of two shapes), using ratios ranging from 1.05 to 3.3. Children were given pre-recorded feedback about the accuracy of their answer.

Confidence Task. We used the Prospective Confidence task from Study 5, but without the warm-up trials, as all children completed the area task first.

Results

Selective Social Learning Task. As shown in Table 4.1, children at all three ages identified the informant with lesser errors as the Winner more often than chance, with a significant increase with age, $r(130) = .43$, $p < .001$. Six- and 7-year-olds also Asked this informant for help, and in contrast with both Study 5 and 6 also Endorsed her shape as being closer, while 5-year-olds did neither (see Table 4.1). Ask choices correlated with age, $r(130) = .45$, $p < .001$, and Endorse choices had trending correlation with age, $r(130) = .17$, $p = .055$. Children's choice of Winner was correlated with who they Asked for help, $r(129) = .55$, $p < .001$, and who they Endorsed, $r(129) = .39$, $p < .001$, and Ask and Endorse responses were also correlated, $r(129) = .40$, $p < .001$.

Area Task. Children in all three age groups successfully identified the larger shape well above chance levels (see Table 4.1), and accuracy did not significantly increase with age, $r(130)$

= .11, p = .213. Performance on the area task was not significantly correlated with children's answers on Winner, $r(130) = .03$, p = .756, Ask, $r(130) = -.02$, p = .805, or Endorse questions, $r(130) = .02$, p = .806.

Confidence Task. Replicating both Studies 5 and 6, children at all ages also chose the easier question more than expected by chance. There was no correlation with area performance, $r(130) = -.01$, p = .939. Children were more likely to choose the easier question on larger metaratios, $F(2, 258) = 6.90$, p = .001, $\eta_p^2 = .05$, but there was no improvement with age, $F(2, 129) = 1.26$, p = .286, $\eta_p^2 = .02$, and no interaction, $F(4, 258) = 0.18$, p = .950, $\eta_p^2 = .00$, replicating Study 5.

Replication of Studies 1 & 2. Age and area discrimination together did not significantly predict children's confidence discrimination, $R^2 = .03$, $F(2, 129) = 2.06$, p = .131, R^2_{Change} from model with just age = .00, $F(1, 129) = 0.09$, p = .766, though the coefficients suggest that age was a more meaningful predictor than area, consistent with what we found in Studies 1, 2, & 6, $\beta_{Age} = .18$, $t(129) = 2.03$, p = .044, $\beta_{Area} = -.03$, $t(129) = -0.3$, p = .770.

Correlations of Individual Differences. As shown in Table 4.2, there were no correlations between the confidence discrimination task and any of the selective social learning measures when controlling for age and area discrimination.

Discussion

We found that both tasks replicated the key patterns from past work, indicating that they were tapping into the target constructs. However, even when using a larger sample size powered to detect the observed effect sizes from Studies 5 and 6, we did not find any correlation between children's sensitivity to confidence and their selective social learning choices.

Mega-Analysis

Given the apparent inconsistencies in these results and the limited availability of the vocabulary assessment, we conducted additional analyses using data combined from the studies to increase power.

Vocabulary Assessment

In Studies 6 and 7, we administered a short vocabulary assessment to act as a coarse control for intelligence, but approximately only 1/3 of parents completed the questionnaire. With these data combined (N = 72, M = 115.68, SD = 33.67), we found that DVAP scores increased with age, $r(70) = .30$, $p = .011$, as expected. When controlling for age, DVAP scores correlated with children's choice of Winner, $r(69) = .40$, $p < .001$, and who they Asked for help, $r(69) =$ ..25, $p = .032$, but not who they Endorsed, $r(69) = .08$, $p = .502$, or their performance on the confidence task, $r(69) = .06$, $p = .648$.

However, no correlations between the confidence measures and the selective social learning measures reached significance when DVAP scores were controlled for along with age and area (see Table 4.2).

Selective Social Learning Task

Across the three studies, we found variability in the age at which children's judgments surpassed chance levels, though in several cases children's performance was consistent with our prediction that children would choose the informant with lesser errors across the three responses. Combining the data from all three studies, we find that children aged 5 and up gave Winner and Ask judgments consistent with our prediction, and children aged 6 and 7 gave consistent Endorse judgments (see Table 4.1). Thus, largely replicating the findings of Einav and Robinson (2010), we found that children were sensitive to the relative degree of error, though we did not find evidence that 4-year-olds identified or strategically trusted relatively more accurate informants.

Correlations of Individual Differences

There was a small but significant correlation between confidence sensitivity and Winner and Ask choices, but not Endorse choices (see Table 4.2). However, none of these correlations held when controlling for age (available for all 3 studies), or for age and area discrimination (available for Studies 6 and 7), suggesting that any correlations were likely being driven by other individual differences.

General Discussion

We set out to test how broadly confidence representations are calculated by looking for correlated individual differences in a relative measurement of own confidence and a relative assessment of others' accuracy. Across three studies, we replicate the key findings of both the selective social learning measure and the relative confidence measure, but do not find that they correlate independently of age and area abilities (the Type 1 task used). Even under a generous interpretation of our findings, there is at most a very small correlation between the tasks, certainly much less than would be expected if the two tasks relied on the same ability to compute confidence. We are therefore most inclined to interpret these results as evidence that reasoning about the strength of one's own knowledge (i.e., confidence) and the strength of another's knowledge are not computed by a single process.

These results are inconsistent with both the mind-reading first account and the self-first/simulation accounts of Mind-Reading theories, which argue that decisions about the reliabilities of self and other knowledge are computed as a single process. In fact, and consistent with our results, a growing body of work suggests that the age-related overlaps in mind-reading and metacognitive abilities are very likely due to shared linguistic markers and response biases (Bernard et al., 2015; Resendes et al., 2019). For instance, when using a metacognitive task that limits language requirements (e.g., an opt-out task, see Balcomb & Gerken, 2008) and a mind-reading task that does not rely on one's own help-seeking biases (e.g., a false belief task, see Wimmer & Perner, 1983), children's self and other reasoning was uncorrelated (Bernard et al., 2015). Similarly, an ongoing study found that differences in reaction time on accurate and inaccurate items (used to assess non-verbal confidence monitoring) did not predict children's selective choice of a previously accurate or inaccurate informant, but the accuracy of their explicit confidence judgments did (Resendes et al., 2019). Together with the present findings, this suggests that links between mind-reading and metacognition may be driven by shared task demands (seeking more information from others and using mental state language) rather than by shared computations of confidence.

At the same time, there could be alternative explanations for the lack of correlations in the current study that stem from specific task parameters we used. First, mind-reading-first accounts propose that children evaluate the likelihood of their own accuracy and others' accuracy by reasoning about behavioural cues (Carruthers, 2009; Gopnik, 1993), which we

significantly limited in our selective learning task by using static photos of the informants rather than dynamic videos or in-person demonstrations. While children in our studies *were* clearly able to use information only about the relative sizes of the shapes to infer the relatively more accurate informant (as demonstrated by their above-chance performance in the social learning task), their decisions in the confidence task could have been informed by much richer behavioural cues (e.g., reaction time, states of anxiety, emotional expression), leading to a dissociation between the two. Second, both tasks we used relied on children's strategic choices to maximize their performance (e.g., choose the more reliable teacher, choose the easier question to answer), which may not necessarily have operated identically across the two tasks. For instance, a child may have been particularly motivated to be 'fair' to the two informants, sacrificing strategic reasoning for a normative goal (e.g., Shaw & Olson, 2012), but still be very strategic in using confidence to maximize their own performance. This child's performance would then appear chance-like on the social learning task, but not on the confidence task (and there could also be children with the opposite pattern – motivated to select the best informants but unsystematic when reasoning about own performance). These remain possibilities for future research to test, possibly with stronger reward structures and in-person contexts.

Despite these potential confounds, our data remain consistent with the predictions of the Bayesian account – that confidence is represented in a common format even if it is computed independently based on the decision at hand. In the current studies, children are never directly required to compare their own internal confidence about a decision with their evaluation of another agent's likely accuracy (the way they had to in Study 4), and thus our current findings neither support nor refute this account. That said, there are many studies that strongly suggest that children *do* compare confidence flexibly between the self and others (Baer & Odic, in preparation; Bridgers et al., 2016; Harris et al., 2018; Jaswal et al., 2014; Magid et al., 2018; Mascaro & Sperber, 2009; Mills, 2013). As one example, children as young as 4 years old strategically give a younger partner easy tasks while themselves taking on difficult ones (but reverse this behaviour when working with an older partner), thereby using a metacognitive evaluation of difficulty with expectations about age-related abilities to compare each party's relative likelihoods of success (Baer & Odic, in preparation; Magid et al., 2018). Similarly, preschool children are selectively skeptical of adults who provide information that conflicts with a child's own knowledge, suggesting that they weigh the relative likelihood of their own

knowledge against incoming evidence (Jaswal et al., 2014; Mascaro & Sperber, 2009; Mills, 2013). While our results speak against the evaluation of these diverse sources through a single domain-general cognitive process, they do allow for these comparisons to occur given they are made with a domain-general *unit* (e.g., the overall probability of an answer's accuracy).

These results also provide important replications of both the selective social learning task and the relative confidence task. Here, we asked children to reason about the relative accuracy of two informants by comparing the relative sizes of objects, rather than numerical estimates or categorical labels as used by Einav & Robinson (2010). We found that children as young as age 5 not only noticed the difference in sizes and attributed this to superior ability (by choosing the closer informant as the 'Winner'), but they also used this attribution to strategically ask for help ("Ask" questions) and by age 6 to reason about ambiguous cases ("Endorse" questions). We can therefore echo their conclusion that children are sensitive to the magnitude of an informant's error and do not reason about accuracy only in binary terms.

We do, however, see some differences in our pattern of results compared to the original findings of Einav & Robinson (2010). Most notably, we did not find that 4-year-olds attributed any differences in the magnitude of error to the informants' ability, nor did they selectively Ask or Endorse the closer informant. It is highly unlikely that this is due to a difficulty in reasoning about area judgments relative to number judgments, as there is ample evidence that reasoning about area is well-developed by this age (and certainly for the ratios used in this task), while numerical reasoning continues to develop into later childhood (Odic, 2018). Instead, we suspect that there are three likely explanations that are not mutually exclusive. First, our task could have been less engaging for children because of the use of pictures rather than videos or live demonstrations, and so these youngest children could have been unmotivated to respond strategically. Second, children were not given a number line to record the answers of each informant, as they were in the Einav & Robinson study, which could mean that 4-year-olds understand the general principle that 'closer is better', but do not spontaneously track the magnitude of error without help. Third, it could also be that 4-year-olds do not possess a "closer is better" rule at all, as 4 and 5-year-olds were considered as a single age group in Einav & Robinson's study, so older children could have driven this effect.

We also found that children in Studies 5 and 6 did not strategically Endorse the closer informant, even though they chose her more often than chance in both Winner and Ask

questions. This changed in Study 7 with a seemingly small difference in the question wording: from “Which one do you think is more like the [target shape]?” to “Which girl’s shape would you guess looks the way it is supposed to look?” This change introduces two possible explanations for differences across studies. One is that by drawing attention to the informants as the owners of the shapes, children may have been more likely to think about the informants and their abilities than without this cue. The second is that by using language that acknowledges the ambiguity and imperfection of the situation (“would you guess” and “supposed to look”), children may have felt more comfortable repeating their choice of informant (e.g., Bonawitz et al., 2020) or using plausible deniability to override a desire to be fair to both informants (e.g., Shaw et al., 2014). Future work testing each of these possibilities could be useful not only in understanding children’s behaviour on selective learning tasks, but more generally in understanding how children balance epistemic and social goals (Jaswal & Kondrad, 2016; Landrum et al., 2015).

We also replicated and extended many of the findings from Studies 1 & 2 in the confidence task. As we saw in Studies 3 & 4, children responded to their confidence in area discriminations by selecting the easier question (Studies 5 & 7) or the more accurate answer (Study 6). Extending this work, we saw that even 4-year-olds were significantly above chance in their confidence reasoning in all three studies, supporting the prediction made in Chapter 2 that children under age 5 may show sensitivity to confidence if given large enough contrasts in difficulty. We also found that there was independent development of confidence in Study 6 once Type 1 area abilities were controlled, replicating the key findings of Studies 1 & 2, though we did not find a strong replication in Study 7 when we used the Prospective Confidence task. It is therefore possible that the confidence stimuli used in Studies 5 & 7 did not give us sufficient variability to detect these age-related correlations. If true, this could also provide an explanation for why we did not see correlations between the confidence task and the social learning task in Study 7. However, this likely does not affect our main conclusion that there is no strong correlation between the two tasks, as even when we *did* find a correlation between age and confidence sensitivity in Study 6, there was still no correlation between confidence sensitivity and selective learning.

Chapter 5: Discussion

In a world filled with confusion, humans need a toolkit of skills to help discern fact from fiction. Throughout this dissertation, we have explored how children reason about the strength of subjective information – their metacognitive reasoning about confidence – uncovering the properties of childhood metacognition and using its development as a tool to learn about metacognition as a whole.

Fueling this research were two families of accounts that make divergent predictions about the information used to calculate confidence judgments, the breadth of confidence judgments, and the potential sources of change with development. At one end, we have Direct accounts that theorize confidence to be a direct readout of perceptual noise (or memory noise, etc.; Baranski & Petrusic, 1998; Pouget et al., 2016; Salles et al., 2016). Given that there is a direct link between the information used to make the decision and information used to judge confidence, these accounts therefore make a strong prediction of highly correlated Type 1 (decision) and Type 2 (confidence) performance. This similarly carried forward to a prediction of domain-specificity in confidence across Type 2 judgments for unrelated Type 1 tasks (e.g., number and emotion). Importantly, developmental change in metacognitive reasoning is argued to come only from two sources: changes in response biases (e.g., a reduction in overconfidence) or changes in Type 1 noise (e.g., developing perceptual systems).

At the other end are Inferential accounts that theorize confidence to be computed from information over and above direct noise readouts, such as behavioural cues (Carruthers, 2009; Koriat, 1993; Meyniel et al., 2015). Inferential accounts can still accommodate a link between Type 1 and Type 2 judgments (e.g., if the cues are reliable indicators of accuracy), but also predict additional variability unique to Type 2 judgments. Confidence judgments could therefore be correlated across domains if the information combined to make a confidence judgment in one domain is the same as that combined in another domain (e.g., reaction time). One type of Inferential account, the Bayesian account, specifically predicts that regardless of whether metacognition correlates across domains, it should universally use a common probability unit (the likelihood of a decision being correct). Developmental change, then could come from improvement unique to the consideration or integration of certain cues, or in the development of a confidence-specific processor.

Across 7 studies, these predictions were put to the test using a novel measure of metacognition for children that limited the need for mental state language and eliminated the influence of response biases. This measure relies on a relative assessment of confidence – selecting which of two items yields higher confidence – that removes the need for children to set a threshold for high versus low confidence (Barthelmé & Mamassian, 2009; Butterfield et al., 1988; Lipowski et al., 2013; Nelson, 1984). In all 7 studies, children reliably chose items they were most likely to get correct, replicating many studies finding metacognitive monitoring in children (Ghetti et al., 2013; Goupil & Kouider, 2019) and demonstrating that this method of obtaining confidence judgments from children is effective as young as 4 years of age. Notably, children made strategic confidence-based decisions both before and after answering the Type 1 questions, and in three distinct perceptual domains (number, area, and emotion), demonstrating versatility both in children's metacognition and in this paradigm.

In three of the six studies with multiple age groups (Studies 1, 2, & 6, but not 3, 5, & 7), older children had more sensitivity to confidence than did younger children. Because overconfidence biases do not impact the relative confidence task, this developmental effect cannot be attributed to changing response biases. Moreover, in three of four studies that measured both Type 1 and Type 2 performance (Studies 1, 2, & 6, but not 7), confidence sensitivity correlated with age even once Type 1 sensitivity was statistically removed. Together, these findings suggest that overconfidence and Type 1 noise are not the only contributors to metacognitive development, contrary to the predictions of the Direct accounts.

With respect to domain-generality of confidence in childhood, children aged 6 and older had correlated performance on number, area, and emotion perception confidence judgments, even though these judgments themselves (i.e., the Type 1 judgments) were uncorrelated (Study 3). Further, when the relative confidence paradigm was adapted to require that children *compare* their confidence between those same three domains, children were flexibly able to compare within and across domains with equivalent ease (Study 4). While these two studies suggest that confidence is centrally processed and uses a common unit, there was little to no correlation between confidence judgments for the self and social learning judgments based on another's relative accuracy (Studies 5-7). This distinction suggests that there are commonalities in the unit and processing of *perceptual* confidence decisions, at least, but that determining self and other confidence is not identical, as proposed by Mind-Reading accounts.

When all these findings are taken together, we see very little support for the Direct accounts, instead favouring Inferential accounts. These findings therefore provide insight into our understanding of metacognitive processes both in children and humans in general: insight into the information used to make confidence judgments, insight into the breadth of confidence computations, insight into the format of confidence representations, and how we might make use of confidence representations in our broader cognition. Below, I discuss each of these in turn.

Where do the Direct and Inferential Accounts Stand?

Under the Direct accounts, confidence represents a direct readout of Type 1 noise (once response biases are properly controlled for). And, in three of these studies (Studies 1, 2, & 6), children's confidence sensitivity correlated with their Type 1 sensitivity (number or area), signalling that the noise of perceptual representations likely contributes to confidence judgments (Halberda & Odic, 2014; Maniscalco & Lau, 2012). Nonetheless, Type 1 noise doesn't *fully* explain development in confidence sensitivity. For one, these correlations were at most medium effects (*r* values ranging from -.01 to .42), certainly not the strong correlations expected by a Direct account. Then, we saw that age remained a strong predictor of confidence sensitivity when number or area precision was controlled for, suggesting independent development of the confidence sense. And, as a strong case that Type 1 noise alone is insufficient to calculate confidence, we also saw in Chapter 4 that two kinds of confidence judgments (self and other) with identical Type 1 noise (area discriminations) did not correlate strongly if at all. While none of these findings should be interpreted as definitive evidence that Type 1 noise is irrelevant for confidence judgments, we can conclude that children likely use additional information to make confidence decisions.

With that said, is there any way the Direct accounts could still explain these findings? Possibly – because the relative task asks children to make a confidence judgment separately from their Type 1 decision, there are two potential sources for noise. From the perspective of the Accumulator account, the same information available to make the Type 1 decision may continue to accumulate over time before making the Type 2 confidence judgment (Baranski & Petrusic, 1998; Pleskac & Busemeyer, 2010). Because this information is noisy, it may occasionally or even frequently cause children's confidence to fluctuate after making the decision. Given that we know Type 1 noise generally decreases with age (Odic, 2018), this post-decision accumulation

would be much noisier for younger children than for older children, which could lead to less accurate confidence choices for younger children relative to older children. From the perspective of the SDT account, which proposes that information is extracted only once to make both Type 1 and Type 2 decisions, there could be decay in the quality of the extracted information over time. This could potentially be related to working memory, which is also known to develop through childhood (Gathercole et al., 2004), leading to the same prediction that younger children (who could experience more decay) would appear to have worse metacognitive abilities than older children. If true, we would predict that confidence judgments made *following* the Type 1 decision would be subject to more age-related change than confidence judgments made *before* the Type 1 decision, as only post-decision confidence is subject to these two explanations.[20] Across our studies, there is some evidence of such a pattern: Studies 1, 2, 3, 5, and 7 all used the Prospective Confidence task (in which children chose the trial they were more confident in *before* giving us the answer) and all but Study 1 found weak or no development once Type 1 noise was controlled (when possible). In contrast, Studies 1, 2, and 6 all used the Retrospective Confidence task (in which children first make the Type 1 decision and *then* judge their confidence) and all found evidence of development even once Type 1 noise was controlled. The Direct accounts could, therefore, account for the dissociations we observed between development and confidence by arguing that children's post-decision confidence is affected by the continued noisy accumulation or the decay of evidence, while their pre-decision confidence better captures their true confidence reasoning.

However, two additional key findings in these studies make this explanation unlikely. First, in both Studies 1 and 2, children's confidence sensitivity correlated between prospective and retrospective tasks even *after* controlling for both age and numerical precision (i.e., Type 1 noise), suggesting that there was meaningful confidence-specific variability in *both* tasks that was not solely dependent on their Type 1 noise (or even dependent on age). Second, the evidence of domain-generality in Chapter 3 points to a process and unit of confidence that exist at least somewhat independently of the Type 1 decision, contrary to the predictions of the Direct accounts. It is therefore highly likely that additional sources of information are available to confidence judgments that are not based on a direct noise readout.

[20] This assumes that children in the prospective tasks did not answer each question in their head before making a confidence judgment, which cannot be ruled out with the given data (see Pouget et al., 2016).

There is still one remaining explanation that a Direct account could use: maybe the remaining development we see is the result of performance demands in the relative task. This task, for all its benefits in reducing the task demands associated with language and interpreting confidence with respect to an arbitrary scale, does still rely on strategic reasoning and a motivation to be accurate. That is, children must want to be accurate and understand (at some level) that choosing the item they feel most confident in will maximize their accuracy, and these could change with age. Of course, this critique is not unique to relative confidence tasks. Participants similarly need to be motivated to accurately report their confidence to an experimenter, or to understand that choosing a high-confidence item will maximize a reward. To combat these concerns in studies with animals and children, some tasks implement a 'payout matrix' by motivating participants with additional rewards if they indicate high confidence when correct or by retracting rewards when confident but incorrect (Kepecs & Mainen, 2012; Vo et al., 2014).

There are several findings in the current data and in related studies that make this explanation unlikely, though. In other work with this relative task, implementing such a reward system did not improve 4-6-year-old children's confidence sensitivity (Baer & Odic, 2020), making it unlikely that motivational differences could explain the age effects (though individual differences certainly exist, as illustrated by the small proportion of children who consistently choose the harder option, which could potentially explain the correlation between prospective and retrospective tasks in Studies 1 and 2)[21]. Furthermore, if differences in motivation or strategy were all that contributed to confidence variability, then we should not have seen metaratio effects – accuracy should have been similar across the metaratios. But perhaps most convincingly, notice that concerns of motivation and strategy use do not apply to Study 4 where children were found to use a common unit across perceptual tasks, as this study used a within-subjects experimental design that controlled for these explanations. There, individual differences in strategy use or motivation would have at most dampened the overall effect (e.g., some children would have been motivated throughout and shown the effect, but others would not). Despite this, we found that children could compare their confidence across perceptual boundaries, something

[21] An important caveat here is that children did not always consistently choose harder questions in both versions of the task. Some children only did so on one task but not the other, but it is worth noting that the results remain nearly identical with these children's data inverted or removed.

that should not be possible if confidence exists in the units of Type 1 decisions (the way we cannot directly compare two standard deviations to one another without first standardizing them into common units). Overall, then, it seems unlikely that a purely Direct account could accommodate the current findings.

The Inferential accounts, in contrast, easily accommodate all the evidence by allowing for additional sources of information to make confidence judgments. Type 1 noise can be incorporated into the decision, but so can other information like reaction time, beliefs about one's ability, or other cues relevant to the decision at hand. This naturally explains why we found remaining development when controlling for Type 1 noise: there are more developing sources of information. To accommodate the integration of these varied sources, we can think of confidence as a probability estimate about the accuracy of a decision (Meyniel et al., 2015; Pouget et al., 2016). Using a common unit like probability, consistent with the domain-general comparison results of Study 4, allows all sources of information to exist on the same scale and through Bayesian inference[22] these could be integrated to form a cohesive and unified sense of confidence.

What Information is Children's Confidence Based On?

If confidence is not a direct noise readout but rather based on additional information, what might this additional information be? Many potential cues to confidence have been proposed over the years (for examples, see Alter & Oppenheimer, 2009; Kahneman & Tversky, 1982; Koriat, 1997; Shah & Oppenheimer, 2008), so for brevity I will discuss only a few here.

One particularly well-studied cue is decision time (or choice latency, reaction time, etc.), which generally correlates with both accuracy and confidence (Pleskac & Busemeyer, 2010; Rahnev et al., 2020)[23]. Children could therefore attend to the time it takes to answer the Type 1 question and integrate their assessment of their likely accuracy given their decision time with similar assessments using Type 1 noise and other cues to determine which answer was easier. The current studies were not designed to test this possibility as all response buttons were pressed by the experimenter (to reduce motor, memory, and task-switching demands on the children),

[22] Or another type of computation that relies on probabilities (Adler & Ma, 2018).

[23] Decision time is a critical component of Accumulator models, as well, but some models have proposed that *only* decision time affects confidence judgments, without direct access to Type 1 noise (e.g., Koriat & Ackerman, 2010a).

which introduces an extra layer of influence on children's performance (see Study 4). Additionally, children's reaction time varied with the difficulty of the trial in Studies 1 and 2, meaning that decision time could have been a reliable cue for children to use (as it covaried with accuracy), but also that it is difficult to parse apart the effects of Type 1 noise from reaction time. Studies that could manipulate reaction time or run sufficient trials to dissociate it from accuracy and difficulty could reveal what kind of independent contribution decision time makes to confidence decisions over and above Type 1 noise.

Another commonly cited cue is item difficulty, the feature manipulated in these (and most other) studies to induce different degrees of confidence. Difficulty is typically described as a property of a stimulus that on the whole reduces the likelihood that this item will be answered correctly by 'the average person', and as such would reflect a meaningful cue to potential accuracy (Gweon et al., 2017; Nicholls, 1980; Nicholls & Miller, 1983)[24]. However, while difficulty can feel like a single property of a stimulus, it is itself a combination of other potential cues (Gweon et al., 2017). For instance, in the current studies, difficulty was manipulated by changing the ratio of dots (or shape sizes), based on past work showing that different ratios lead to different accuracy (Dehaene, 2011). In other studies, difficulty could be manipulated by adjusting the visibility of a stimulus (e.g., Lyons & Ghetti, 2011), the number of items present (Gweon et al., 2017), or the number of repetitions of an item to be remembered (Hembacher & Ghetti, 2014). Thus, an item's "difficulty" is entirely indexed by the specific objective properties that tend to correlate with accuracy *in that task* (i.e., ratio, visibility, number, and repetition, respectively).

To examine the cue of 'difficulty', then, we must look at the specific objective properties manipulated in each task. In fact, we have already looked at this: reasoning about ratio (the index of accuracy in these tasks) was captured in children's numerical and area comparisons (the Type 1 performance controlled for in Studies 1, 2, 6, and 7; see Jacob et al., 2012), and there was still remaining development. That is, in many cases objective difficulty *is directly reflected* in Type 1

[24] Note that 'difficulty' here refers to the objective properties of the stimulus and not the subjective interpretation of whether an item will likely lead to accuracy. These latter 'feelings of difficulty' are in fact the same thing as feelings of confidence, just with a different focus. As a parallel, we can describe a single scene of two agents running in the same direction both as 'a cop chasing a robber' and as 'a robber fleeing a cop' by shifting the focus from the cop as the main agent to the robber as the main agent – though the scene itself remains the same. Feelings of difficulty are likewise the same feelings of confidence, just directed at the stimulus ('*it* was hard for me') rather than one's own ability to answer it ('*I* was unsure about it').

noise. Nonetheless, difficulty, like confidence, can reflect a combination of objective properties: building a block tower with an alternating colour pattern *and* tiny blocks is harder than building a block tower with only one of these properties (see Gweon et al., 2017). In the case of number representations, we could similarly argue that number judgments are additionally influenced by other non-numeric but magnitude-relevant properties, like the size of the dots or the density of the dot cloud, which have been shown to influence numerical judgments (Clayton et al., 2015; Gebuis & Reynvoet, 2012). While the numerical judgments we obtained from children likely already reflect the influence of these properties, it may still be possible that children rely more heavily on these cues to make their *confidence* judgments, thereby making our Type 1 control insufficient to capture the effects of these objective properties. However, we experimentally controlled for these individual cues to number in Study 2, and it did not result in major changes to children's confidence performance relative to Study 1, making it again difficult to know whether or how much these properties uniquely contribute to confidence judgments. Nonetheless, the current studies did not set out to precisely measure the influence of each of these cues, so future experimental designs could provide more insight by systematically varying objective properties in a way that does not impact accuracy (e.g., by using visual illusions).

Though the current studies were not designed to test the precise effects of different cues, they do provide us reason to believe that an *individual* cue like reaction time or ratio is not solely responsible for children's confidence judgments, the same way we saw that Type 1 noise cannot fully explain confidence sensitivity. More likely, then, is that many cues together lead to a unified sense of confidence, as predicted by most Inferential accounts (Carruthers, 2009; Pouget et al., 2016). These cues could of course include decision time or indexes of difficulty, but perhaps more interestingly potential cues may be limited only by the subject's own beliefs about what information predicts accuracy. For example, having a history of accuracy on a given task could generally increase confidence on that task in the future, leading a subject to rely on broader self-beliefs about their ability rather than moment-to-moment fluctuations in task performance (e.g., Odic et al., 2014). Or perhaps every time the child answers correctly, a nearby adult gives them a thumbs-up, leading the child to rely on adult encouragement as a cue to their accuracy (e.g., Selmeczy & Ghetti, 2019). So long as a cue is deemed relevant for accuracy and can be interpreted as reflecting a given probability of accuracy, the Inferential accounts can accommodate it as a cue for confidence. But, this does not yet explain why we see metacognitive

development. If we permit that there can by any number of potential cues, how can we account for systematic improvement in metacognitive abilities in using these cues?

Sources of Metacognitive Development

The current studies found that two known sources of developmental change (response biases and Type 1 noise) could not fully account for developmental effects found in past work (e.g., Hembacher & Ghetti, 2014; Lyons & Ghetti, 2011; O'Leary & Sloutsky, 2017; Vo et al., 2014). But in addition to these developing abilities, there are at least two non-mutually exclusive ways in which we could explain developmental change: (1) children might need to learn which cues are relevant (and *how* relevant each cue is); and (2) children may already possess (at least some) cues to confidence, but fine-tune their confidence representations over time.

First, there could be an infinite number of relevant cues for the accuracy of a given decision, and so one critical source of development might be isolating the relevant cues for each given decision. For instance, recall the 'sticky mittens' experiment described in Chapter 4: infants only attributed goals to a hand that consistently grasped one of two objects if the infants were first given the (novel) opportunity to grasp desired objects with their own hands (aided by sticky mittens; Sommerville et al., 2005). There, we could describe infants as unaware that grasping was a cue to goals until they were first able to link their *own* goals to a grasping-like action. In a similar way, we could argue that many cues to confidence must be learned over time, say by observing accurate outcomes following short decisions to learn that reaction time is a relevant cue (Koriat & Ackerman, 2010a). Then, even if we granted children knowledge of relevant cues, there could still be change in how such cues are integrated to form a single confidence judgment. For example, children may have to learn which cues are *most* relevant for a given decision, which could rely on developing working memory capacity, attentional resources, or effective connections between cortical regions. In support of a "cue-learning" account, we could interpret the differences in remaining development in the retrospective tasks (Studies 1, 2, & 6) relative to the prospective tasks (Studies 1, 2, 3, 5, & 7) as evidence that the additional influence of cues on the retrospective task led to increased developmental change.

Second, the findings of Study 4 suggest that confidence could be represented in a format akin to probability (a continuous dimension), which opens the possibility that sensitivity to probability itself develops. For example, we know that number similarly exists on a continuous

dimension, and one of the key forces of development in numerical reasoning is the precision with which we can tell apart points on that dimension – from very large ratios like 2:3 in infancy to much smaller ones like 10:11 by adulthood (Halberda et al., 2012; Halberda & Feigenson, 2008; Izard et al., 2009). An intriguing possibility is that confidence similarly exists on a graded continuum which can be discriminated more precisely with age. Accordingly, across five of the 6 studies that examined metaratio effects (the ratio between Type 1 ratios, in Studies 1, 2, 3, 6, & 7, but not 5), children were much more likely to indicate high confidence in larger (easier) ratios than smaller ones, something that has been demonstrated in adult participants as well (Barthelmé & Mamassian, 2009; De Gardelle & Mamassian, 2014). Within SDT frameworks, ratio-dependency is considered to be a consequence of perceptual noise, which can change over time as cognition develops and declines (e.g., Halberda et al., 2012).[25] Therefore, we could similarly reason that with age, children also become better attuned to evaluating and comparing *confidence noise* (i.e., probability), discerning even closer differences as they develop.

There are several options available to test these possibilities in future work. For one, studying children at younger ages could help test between these explanations. Children much younger than age 4 have responded metacognitively in other measures (Goupil et al., 2016; Goupil & Kouider, 2016), making it possible that they could use metacognitive reasoning in a relative paradigm as well. If cue-learning is responsible for development, then we might expect younger children to perform at chance if they lack the relevant cues. However, if sensitivity to probability is developing, then younger children might simply need easier metaratios to discriminate (and see Chapter 4 for preliminary evidence of this). Another avenue is to look for cross-cultural and other systematic environmental differences. Cross-cultural data is currently very rare in the study of confidence, but there are known differences in the degree to which certain languages talk about confidence. For instance, Turkish speakers provide evidential markers to indicate the source of claims (e.g., direct/indirect experience, second-hand report; Aksu-Koç et al., 2009), which could mean that Turkish children are forced to attend more to

[25] Because we manipulated ratio to induce different levels of confidence, we could only capture differences as they related to ratio. However, the actual dimension is likely not based on ratio (as discussed earlier), but more likely probability of success or a similar domain-neutral quantity. Accordingly, identical metaratios instantiated with different ratio pairs (e.g., a metaratio of 2.0 made with ratios 4.0 and 2.0 could also be made with ratio 2.1 and 1.05) may not lead to identical discriminability if the participant's actual probability of success was not equated between these pairs (e.g., the first pair might be a difference of 99% vs. 99.9% correct, whereas the second might be 75% vs 99%, two drastically different comparisons).

certainty-relevant information in their daily lives. Similarly, many teachers and professors preach the importance of developing critical thinking skills in their students, which relies on evaluations of confidence. In both cases, we might expect to find differences in the effectiveness of cue use or in the precision of probability reasoning. Lastly, longitudinal designs could be valuable in tracking changes over time within the same individual, rather than cross-sectionally. In combination with targeted training on probability reasoning or attention to cues, longitudinal designs could test how cues emerge over time and how certain experiences lead to metacognitive change.

Is Confidence Computed Domain-Generally?

With the possibility of additional information contributing to confidence judgments, we should then wonder whether there is a central system for processing confidence, or whether confidence judgments still retain some domain-specificity. We saw in Chapter 4 that even when using the same Type 1 task, reasoning about self and other confidence was not strongly related. This simultaneously provides evidence that Type 1 performance is insufficient to explain metacognitive abilities and that self-focused confidence reasoning does not extend so far as to include other-focused confidence evaluation, in contrast with both a strong Direct account and some Inferential accounts (e.g., Mind-Reading). But then in Study 3, children's confidence sensitivity *was* correlated in three perceptual domains that were unrelated at the Type 1 level, strongly suggesting that metacognitive processes were shared among the three domains. How broadly, then, is confidence computed?

The findings of Study 3 stand in contrast to one other study with children, which found no correlation between number and emotion metacognition (two domains that were found here to correlate), but are consistent with much of the work in adults showing domain-general perceptual metacognitive processes (Rouault et al., 2018; Vaccaro & Fleming, 2018; Vo et al., 2014). From this, we could make a few conclusions: maybe Vo and colleagues (2014) failed to capture a true correlation, or maybe the correlations in Study 3 were due to correlated motivation, strategy use, intelligence, or other third variables we did not measure. These are all certainly possible, and could be tested with replications that control for third variables (or experimental designs like Study 4, which suggest that these third variables are not enough to explain domain-generality).

However, there is another possibility worth considering that accommodates both findings as-is. It could be that *sensitivity* to confidence is domain-general, that there is a single continuous sense of confidence used across domains, but our *biases* to report or use this confidence are domain-specific (see Baer & Odic, 2020). As one example, we might find that children's beliefs about their ability in each domain are unrelated – a child could feel skilled at the number task because they are good at math, but terrible at the emotion task because they don't engage with others often. Regardless of whether these self-assessments are true, these beliefs could impact where the child sets their confidence threshold, such that it is different in each task. Domain-generality of sensitivity but not bias is similar to the 'tagging' hypothesis recently put forward to explain why perceptual metacognition appears to be widely correlated in adults, but memory and perceptual metacognition are not consistently related (Rouault et al., 2018). Under this hypothesis, confidence is computed using a single cognitive resource, then tagged with domain-specific information in order to make it useful to decision-making. If true, we might expect that these top-down self beliefs would only affect performance in absolute tasks (where the child rates their confidence or must decide to act or not) rather than the relative task used here or other tasks that separate sensitivity from bias.

Another option is that that confidence is computed domain-generally, but that different cues are used to make confidence judgments in different domains. Consider a simple case in which perceptual confidence is the result of Type 1 noise combined with decision latency. Though sensitivity to Type 1 noise will be independent for each perceptual decision a subject makes, sensitivity to decision latency would be common to all perceptual decisions and could therefore be responsible for metacognitive correlations between domains. Then, imagine that determining another's confidence relies only on sensitivity to facial expression and not decision latency. Though we could have the exact same process compute confidence as for perceptual decisions, sensitivity to another's confidence might not correlate with perceptual confidence because it is subject to the sensitivity of a different set of cues (see Lee et al., 2018). This could potentially explain why perceptual and memory metacognition are often uncorrelated in adults (Baird et al., 2013; Rouault et al., 2018; Vaccaro & Fleming, 2018) – confidence decisions in different cognitive domains could make use of different cues each with their own uncorrelated variability.

Yet another possibility is that there is metacognitive domain-specificity, but that which is still separate from Type 1 processing. For instance, we might find that different decisions are processed in a similar way by one 'company,' but in different 'departments.' Metacognition would thus be domain-specific because each department operates semi-independently, but we could see aspects of domain-generality because of broader 'company policies' (like shared cues, a shared representational format, etc.), and because each department takes on metacognitive decisions for broad categories of cognition (e.g., all perceptual judgments, like the domain-generality found in the current work).[26] This could accommodate the tagging hypothesis (departments follow standardized protocols, but within their respective domains of expertise), and also differences in cues (each department uses a different 'supply chain'). We could also accommodate development under this view as departments 'merge', communicate more effectively, or share their supply chains.

Developmental investigations would be particularly helpful in teasing apart these options. For instance, a tagging hypothesis presents domain-generality as the default, which should mean that if there is any developmental change, it should emerge in the direction of domain-generality to domain-specificity as the mind learns how to effectively tag the common confidence signal. In contrast, a 'departmental' account would predict the opposite: early domain-specificity merges to domain-generality over time. As an intermediate account, an account that proposes domain-generality but with domain-specific cues would predict a mix of effects, including early domain-generality whenever confidence cues are shared, but equally early domain-specificity whenever they are not.[27] Understanding if and how domain-generality changes over time would narrow down these options.

The current trend reported by developmental researchers is in the direction of domain-specificity to domain-generality. In two recent studies, children's metacognitive abilities were correlated between math, spelling, and memory Type 1 tasks above the age of 8, but *not* at younger ages (Bellon et al., 2020; Geurten et al., 2018). Because both studies used absolute rating tasks (e.g., a 3-point scale), though, we do not yet know whether this switch occurs for

[26] 'Departments' don't need to be categorized into perception and memory and other such cognitive domains. They could instead be categorized into types of confidence judgments (e.g., retrospective confidence judgments, feelings of knowing, judgments of learning, etc.) or even formats of Type 1 tasks (e.g., yes/no judgments or 2AFC tasks).

[27] In combination with a cue-learning account discussed earlier, this could mean that metacognition is largely domain-specific early on as children slowly learn what cues are relevant in each situation.

bias or for sensitivity. The results of Studies 3 and 4, though, suggest that sensitivity for *perceptual* confidence is domain-general at age 6,[28] as it is in adults (Rouault et al., 2018). Interestingly, in ongoing work using a 3-point confidence response scale in combination with the latest statistical measures to separate bias from sensitivity, I am finding domain-*specificity* in children's perceptual and memory metacognition under the age of 8 (Baer et al., in preparation). This is broadly consistent with a departmental or cue-specific account (as the memory and perceptual tasks could rely on different cues), though data collection in still ongoing.

Does Confidence Use a Common Unit?

While the various Inferential accounts are somewhat agnostic to the domain-generality of confidence *processing*, some do make strong predictions about the domain-generality of the confidence *unit*. From Study 4, we saw evidence for a domain-general unit in childhood among three perceptual domains, parallel to work with adults (De Gardelle et al., 2016; De Gardelle & Mamassian, 2014). In ongoing work, I am similarly finding that children can compare their memory and perceptual confidence (Baer et al., in preparation), and compare their own confidence to another's (Baer & Odic, in preparation). These findings, in combination with other findings of flexible self and other confidence comparison (Bridgers et al., 2016; Harris et al., 2018; Jaswal et al., 2014; Magid et al., 2018; Mascaro & Sperber, 2009; Mills, 2013) strongly suggest that confidence shares a broad domain-general unit despite differences in cognitive processing.

A domain-general unit stands in contrast to Direct accounts of confidence which predict that confidence exists in the same units as the Type 1 decision (like the standard deviation of a sample, which is only meaningful if you know the relevant units).[29] However, the Bayesian account assumes that confidence is represented probabilistically, in a common unit reflecting the likelihood of accuracy (Meyniel et al., 2015; Pouget et al., 2016). So long as every confidence judgment is made about the accuracy of a decision, this allows for flexible combination and comparison across diverse cues and Type 1 tasks. A probabilistic unit also helps explain the

[28] It is possible that we did not capture early perceptual domain-specificity either because of low power to detect age effects (10 children per age group) or because the shift within perception occurs younger than age 6. Studies with younger children will be required to test this possibility.

[29] Nonetheless, this view could account for Study 4 by arguing that by age 6, children become able to translate between domain-specific units, much like a currency exchange does with two independent currencies. Again, this would require testing for a domain-general unit in younger children.

presence of metaratio effects we saw in many studies, reflecting reasoning about a continuous and noisy scale. Accordingly, we see many parallels in children's reasoning about *objective* probabilities as in subjective probabilities like confidence. From age 7 to age 12, children's probability reasoning not only developed, but followed metaratio effects like those shown here (O'Grady & Xu, in press). And, much like we saw evidence of metacognitive reasoning in infancy (Goupil et al., 2016; Goupil & Kouider, 2016), there is ample evidence of infants performing simple probabilistic calculations like which of two rewards is more likely from a given distribution (Denison et al., 2013; Denison & Xu, 2019; Kayhan et al., 2018). We should therefore expect to see links between the development of probabilistic reasoning and metacognition.

What is the Impact of Reasoning About Confidence?

As with any cognitive or even biological process, we ultimately experience confidence because it is useful to us. For instance, this work signals that confidence may exist as a common unit between decisions from vastly different cognitive domains even in childhood (De Gardelle et al., 2016; De Gardelle & Mamassian, 2014), which provides a mechanism by which domain-specific representations could be compared and integrated. As one example, some theorists argue that metacognitive reasoning is the hallmark of conscious experience – the very ability that allows us to realize that we are thinking (and therefore 'are'; Dehaene et al., 2017). Evidence of a common currency greatly aids this account, as one prominent view of consciousness argues that the presence of a global workspace facilitates our conscious thought (Baars, 1993), but does not have a concrete proposal for how domain-specific units could be effectively compared and integrated. If confidence has such a common unit, and exists for all kinds of decisions, then this would be a likely candidate for how conscious thought is achieved (Shea & Frith, 2019).[30]

Within development specifically, several theories have argued that perceptual domains may only be translatable into a common format through the use of language (Carruthers, 2002; Spelke, 2003). For example, young children do not appear to integrate information about the geometry of a room with its visual features (such as colour) unless this information is highlighted with language (Hermer & Spelke, 1996; Hermer-Vazquez et al., 1999). The current findings, on

[30] This account notably does *not* believe that metacognition must be conscious – quite the opposite. Metacognition exists regardless of conscious thought, but is the critical ingredient in facilitating consciousness.

the other hand, suggest that confidence is represented in a format that is easily translatable across independent perceptual boundaries, allowing children to compare which information is more reliable and should, therefore, be used. If true, we would expect that even pre-linguistic children (and even non-linguistic animals), could compare their confidence across modular domains, providing an account for how unified and centralized cognition could be informed by encapsulated perceptual analyzers.

Other work has examined how metacognition is related to broader academic achievement. In many studies, accurate confidence in mathematical answers correlates with math performance on standardized measures (Bellon et al., 2019, 2020; Rinne & Mazzocco, 2014; Vo et al., 2014). However, this correlation doesn't hold when only examining metacognitive *sensitivity* (Baer & Odic, 2020), raising the possibility that the impact of metacognition on learning is driven by the domain-specific biases that allow us to interpret and use our sense of confidence.[31] Accordingly, recent work has found reliable correlations between participants' error detection in numerical tasks (a process that likely involves determining whether confidence in an answer is below a threshold) and mathematical performance (Wong & Odic, 2020).

Reasoning about confidence may then affect the learning process by directing attention; classic studies of metacognition find that school-aged children spend more time studying difficult items than easy ones, a sensible strategy to maximize learning and future success (Lockl & Schneider, 2004).[32] But the benefits of confidence in directing attention likely emerge much earlier than school age. A now-classic paradigm within the study of infant cognition is the *violation of expectation* design. In it, infants are required to either learn or already know a state of the world (for instance, that trucks do not go through walls), and then are shown an event that violates this state (a truck going through a wall). The critical factor in these designs is that infants will naturally be confused about the violation and look longer at the event (Baillargeon et al., 1985), a behaviour that is consistent with gathering additional information to make sense of an

[31] 'Biases' in this sense are not necessarily negative, the way this word is often used. As mentioned earlier, there could be highly adaptive biases to try information slightly harder than one is capable of.

[32] In the current studies, children generally chose to answer easier questions to maximize their current success, though a small subset of children in each experiment consistently chose the harder questions. As noted earlier, this still critically depends on a metacognitive ability to discern which items are likely to lead to success, but with a different mindset toward whether current success or later success is the priority. Interestingly, a parallel literature on achievement motivation finds that while some children adopt an orientation towards mastering difficult items, other children focus on tackling easier items they know they can accomplish (Dweck, 1986; Dweck & Leggett, 1988).

unusual event. Recent studies have supported this interpretation: 11-month-old infants will selectively test relevant explanations for the violation (e.g., hitting a truck to see if it is solid; Stahl & Feigenson, 2015) and are more likely to learn properties of the objects involved (e.g., learning the word 'truck'; Stahl & Feigenson, 2017), suggesting that metacognition is helpful for guiding attention to events that could maximize learning even in infancy (see also Kidd et al., 2012; Vaish et al., 2008). This naturally parallels work with older infants showing metacognitive help-seeking, another kind of information-seeking behaviour. Furthermore, these studies provide evidence that humans, even in infancy, have a complex toolkit for using metacognitive information to guide their learning. The relative confidence paradigm used here could be valuable to the investigation of infants in allowing us to tease apart the contributions of sensitivity to confidence and interpretive biases that allow this information to be used in learning.

Conclusions

We are constantly being bombarded with information and must rely on cognitive tools to sort out fact from fiction. This dissertation explored one such tool that represents the strength of subjective evidence: our sense of confidence. Building on work demonstrating that even children can metacognitively reason about the likelihood that their own knowledge is correct, the current work demonstrates an important place for developmental work within the broader research on metacognition, teasing apart accounts of how confidence is computed. In accordance with recent theories that confidence is a probabilistic representation of the accuracy of a decision, the current evidence of a domain-general unit that greatly resembles a probability judgment also gives additional weight to arguments of rationality in childhood (e.g., Gopnik & Bonawitz, 2015; Sobel & Kushnir, 2013), providing a potential means by which subjective and objective information could be integrated. Far from being subject to the whims of others, children possess a sense of confidence that combines multiple sources in information to create broadly-usable assessments of truth in the world.

References

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology, 14*, e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386. https://doi.org/10.1016/j.cognition.2015.10.006

Aksu-Koç, A., Ögel-Balaban, H., & Alp, İ. E. (2009). Evidentials and source knowledge in Turkish. *New Directions for Child and Adolescent Development*, *2009*(125), 13–28. https://doi.org/10.1002/cd.247

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*, 219–235. https://doi.org/10.1177/1088868309341564

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Baer, C., & Friedman, O. (2018). Fitting the message to the listener: Children selectively mention general and specific facts. *Child Development*, *89*, 461–475. https://doi.org/10.1111/cdev.12751

Baer, C., Ghetti, S., & Odic, D. (in preparation). *Perceptual and memory metacognition in children*.

Baer, C., & Odic, D. (in preparation). *Team players: Children cooperatively adjust the difficulty of problems to their partner's skill level*.

Baer, C., & Odic, D. (2020). The relationship between children's approximate number certainty and symbolic mathematics. *Journal of Numerical Cognition*, *6*.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*, 191–208. https://doi.org/10.1016/0010-0277(85)90008-3

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *Journal of Neuroscience*, *33*, 16657–16665. https://doi.org/10.1523/JNEUROSCI.0786-13.2013

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

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 929–945. https://doi.org/10.1037/0096-1523.24.3.929

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, *21*, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

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

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences*, *107*, 20834–20839. https://doi.org/10.1073/pnas.1007704107

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

Bellon, E., Fias, W., & Smedt, B. D. (2020). Metacognition across domains: Is the association between arithmetic and metacognitive monitoring domain-specific? *PLOS ONE, 15*, e0229932. https://doi.org/10.1371/journal.pone.0229932

Beran, M. J., Decker, S., Schwartz, A., & Smith, J. D. (2012). Uncertainty monitoring by young children in a computerized task. *Scientifica, 2012*, 692890. https://doi.org/10.6064/2012/692890

Bernard, S., Proust, J., & Clément, F. (2015). Procedural metacognition and false belief understanding in 3-to 5-year-old children. *PloS One, 10*, e0141321. https://doi.org/10.1371/journal.pone.0141321

Birch, S. A. J., Brosseau-Liard, P. E., Haddock, T., & Ghrear, S. E. (2017). A 'curse of knowledge' in the absence of knowledge? People misattribute fluency when judging how common knowledge is among their peers. *Cognition, 166*, 447–458. https://doi.org/10.1016/j.cognition.2017.04.015

Birch, S. A. J., Vauthier, S. A., & Bloom, P. (2008). Three-and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition, 107*, 1018–1034. https://doi.org/10.1016/j.cognition.2007.12.008

Bonawitz, E., Shafto, P., Yu, Y., Gonzalez, A., & Bridgers, S. (2020). Children change their answers in response to neutral follow-up questions by a knowledgeable asker. *Cognitive Science, 44*, e12811. https://doi.org/10.1111/cogs.12811

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

Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., & Gopnik, A. (2016). Children's causal inferences from conflicting testimony and observations. *Developmental Psychology, 52,* 9–18. https://doi.org/10.1037/a0039830

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

Cantlon, J. F., Safford, K. E., & Brannon, E. M. (2010). Spontaneous analog number representations in 3-year-old children. *Developmental Science, 13,* 289–297. https://doi.org/10.1111/j.1467-7687.2009.00887.x

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General, 148,* 51–64. https://doi.org/10.1037/xge0000505

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences, 25,* 657–674. https://doi.org/10.1017/S0140525X02000122

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32,* 121–138. https://doi.org/10.1017/S0140525X09000545

Clarke, F. R., Birdsall, T. G., & Tanner, W. P. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America*, *31*, 629–630. https://doi.org/10.1121/1.1907764

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

Coughlin, C., Hembacher, E., Lyons, K. E., & Ghetti, S. (2015). Introspection on uncertainty and judicious help-seeking during the preschool years. *Developmental Science*, *18*, 957–971. https://doi.org/10.1111/desc.12271

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*, 148–153. https://doi.org/10.1016/j.tics.2009.01.005

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

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*, 105–110. https://doi.org/10.1038/nn.3279

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

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*, 486–492. https://doi.org/10.1126/science.aan8871

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: Evidence from 4.5-and 6-month-olds. *Developmental Psychology, 49,* 243–249. https://doi.org/10.1037/a0028278

Denison, S., & Xu, F. (2019). Infant statisticians: The origins of reasoning under uncertainty. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 14,* 499–509. https://doi.org/10.1177/1745691619847201

Dunlosky, J., & Bjork, R. A. (2008). *Handbook of metamemory and memory*. Taylor & Francis.

Einav, S., & Robinson, E. J. (2010). Children's sensitivity to error magnitude when evaluating informants. *Cognitive Development, 25,* 218–232. https://doi.org/10.1016/j.cogdev.2010.04.002

Elmose, M., & Happé, F. (2014). Being aware of own performance: How accurately do children with autism spectrum disorder judge own memory performance? *Autism Research: Official Journal of the International Society for Autism Research, 7,* 712–719. https://doi.org/10.1002/aur.1421

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433. https://doi.org/10.1038/415429a

Filevich, E., Forlim, C. G., Fehrman, C., Forster, C., Paulus, M., Shing, Y. L., & Kühn, S. (2020). I know that I know nothing: Cortical thickness and functional connectivity underlying meta-ignorance ability in pre-schoolers. *Developmental Cognitive Neuroscience, 41,* 100738. https://doi.org/10.1016/j.dcn.2019.100738

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*, 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *1*, nix007. https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Fuster, J. (2015). *The prefrontal cortex*. Academic 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

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*, 177–190. https://doi.org/10.1037/0012-1649.40.2.177

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

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Geurten, M., Meulemans, T., & Lemaire, P. (2018). From domain-specific to domain-general? The developmental path of metacognition for strategy selection. *Cognitive Development, 48*, 62–81. https://doi.org/10.1016/j.cogdev.2018.08.002

Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling uncertain and acting on it during the preschool years: A metacognitive approach. *Child Development Perspectives, 7*, 160–165. https://doi.org/10.1111/cdep.12035

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences, 7,* 43–48. https://doi.org/10.1016/S1364-6613(02)00009-8

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16,* 1–14. https://doi.org/10.1017/S0140525X00028636

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59,* 26–37. https://doi.org/10.2307/1130386

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews. Cognitive Science, 6,* 75–86. https://doi.org/10.1002/wcs.1330

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 3–32. https://doi.org/10.1037/0033-295X.111.1.3

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, *8*, 371–377. https://doi.org/10.1016/j.tics.2004.06.005

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., & Kouider, S. (2019). Developing a reflective mind: From core metacognition to explicit self-reflection. *Current Directions in Psychological Science*, *28*, 403–408. https://doi.org/10.1177/0963721419848672

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 & Sons, Inc.

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 458–463).

Haddock, T., Ghrear, S. E., Brosseau-Liard, P., Baer, C., & Birch, S. A. J. (in preparation). *Fluency misattribution in adults and children's judgements of their peers' knowledge*.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (7th ed.). Pearson.

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

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, *69*, 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Hembacher, E., & Ghetti, S. (2013). How to bet on a memory: Developmental linkages between subjective recollection and decision making. *Journal of Experimental Child Psychology*, *115*, 436–452. https://doi.org/10.1016/j.jecp.2013.03.010

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

Hermer, L., & Spelke, E. (1996). Modularity and development: The case of spatial reorientation. *Cognition, 61,* 195–232. https://doi.org/10.1016/S0010-0277(96)00714-7

Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-Task studies of space and language. *Cognitive Psychology, 39,* 3–36. https://doi.org/10.1006/cogp.1998.0713

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

Jaswal, V. K. (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. *Cognitive Psychology, 61,* 248–272.

Jaswal, V. K., & Kondrad, R. L. (2016). Why children are not always epistemically vigilant: Cognitive limits and social considerations. *Child Development Perspectives, 10,* 240–244. https://doi.org/10.1111/cdep.12187

Jaswal, V. K., Pérez-Edgar, K., Kondrad, R. L., Palmquist, C. M., Cole, C. A., & Cole, C. E. (2014). Can't stop believing: Inhibitory control and resistance to misleading testimony. *Developmental Science*, *17*, 965–976. https://doi.org/10.1111/desc.12187

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11*, 143–157. https://doi.org/10.1016/0010-0277(82)90023-3

Kantner, J., Solinger, L. A., Grybinas, D., & Dobbins, I. G. (2019). Confidence carryover during interleaved memory and perception judgments. *Memory & Cognition*, *47*, 195–211. https://doi.org/10.3758/s13421-018-0859-8

Kayhan, E., Gredebäck, G., & Lindskog, M. (2018). Infants distinguish between two events based on their relative likelihood. *Child Development, 89*, e507–e519. https://doi.org/10.1111/cdev.12970

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*, 92–107. https://doi.org/10.3758/BF03211579

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*, 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*, 227–231. https://doi.org/10.1038/nature07200

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*, 759–764. https://doi.org/10.1126/science.1169405

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, *126*, 109–114. https://doi.org/10.1016/j.cognition.2012.08.004

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks Effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One*, *7*, e36399. https://doi.org/10.1371/journal.pone.0036399

Kikyo, H., Ohki, K., & Miyashita, Y. (2002). Neural correlates for feeling-of-knowing: An fMRI parametric analysis. *Neuron*, *36*, 177–186. https://doi.org/10.1016/S0896-6273(02)00939-X

Kim, G., & Kwak, K. (2011). Uncertainty matters: Impact of stimulus ambiguity on infant social referencing. *Infant and Child Development*, *20*, 449–463. https://doi.org/10.1002/icd.708

Koenig, M. A., & Harris, P. L. (2005a). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*, 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Koenig, M. A., & Harris, P. L. (2005b). The role of social cognition in early trust. *Trends in Cognitive Sciences*, *9*, 457–459. https://doi.org/10.1016/j.tics.2005.08.006

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology*, *52*, 31–45. https://doi.org/10.1037/dev0000065

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

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119,* 80–113. https://doi.org/10.1037/a0025648

Koriat, A., & Ackerman, R. (2010a). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science, 13,* 441–453. https://doi.org/10.1111/j.1467-7687.2009.00907.x

Koriat, A., & Ackerman, R. (2010b). Metacognition and mindreading: Judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition, 19,* 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development, 24,* 169–182. https://doi.org/10.1016/j.cogdev.2009.01.001

Koriat, A., & Adiv, S. (2016). The self-consistency theory of subjective confidence. In *The Oxford Handbook of Metamemory* (pp. 127–147). Oxford University Press.

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 34–53. https://doi.org/10.1037/0278-7393.27.1.34

Kuzyk, O., Grossman, S., & Poulin-Dubois, D. (2020). Knowing who knows: Metacognitive and causal learning abilities guide infants' selective social learning. *Developmental Science, 23,* e12904. https://doi.org/10.1111/desc.12904

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences, 19,* 109–111. https://doi.org/10.1016/j.tics.2014.12.007

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences, 103*, 18763–18768. https://doi.org/10.1073/pnas.0607716103

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

Lee, A. L. F., Ruby, E., Giles, N., & Lau, H. (2018). Cross-domain association in metacognitive efficiency depends on first-order task types. *Frontiers in Psychology, 9*, 2464. https://doi.org/10.3389/fpsyg.2018.02464

Libertus, M. E., Odic, D., Feigenson, L., & Halberda, J. (2015). A Developmental Vocabulary Assessment for Parents (DVAP): Validating parental report of vocabulary size in 2- to 7-year-old children. *Journal of Cognition and Development, 16*, 442–454. https://doi.org/10.1080/15248372.2013.835312

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

Lockhart, K. L., Goddu, M. K., & Keil, F. C. (2017). Overoptimism about future knowledge: Early arrogance? *The Journal of Positive Psychology, 12*, 36–46. https://doi.org/10.1080/17439760.2016.1167939

Lockl, K., & Schneider, W. (2004). The effects of incentives and instructions on children's allocation of study time. *European Journal of Developmental Psychology, 1*, 153–169. https://doi.org/10.1080/17405620444000085

Lyons, K. E., & Ghetti, S. (2010). Metacognitive development in early childhood: New questions about old assumptions. In *Trends and Prospects in Metacognition Research* (pp. 259–278). Springer. https://doi.org/10.1007/978-1-4419-6546-2_12

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

Lyons, K. E., & Ghetti, S. (2013). I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Development*, *84*, 726–736. https://doi.org/10.1111/cdev.12004

Magid, R. W., DePascale, M., & Schulz, L. E. (2018). Four- and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. *Open Mind*, *2*, 72–85. https://doi.org/10.1162/opmi_a_00019

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. In *The Cognitive Neuroscience of Metacognition* (pp. 25–66). Springer. 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*, *2*, 47–60. https://doi.org/10.1162/opmi_a_00017

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*, 367–380. https://doi.org/10.1016/j.cognition.2009.05.012

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

Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000746

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., Lange, F. P. de, & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*, 1897–1906. https://doi.org/10.1523/JNEUROSCI.1890-12.2013

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 851–861. https://doi.org/10.1037/0278-7393.19.4.851

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*, 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology, 49,* 404–418. https://doi.org/10.1037/a0029500

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-General and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience, 38,* 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018

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

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26,* 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

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

Nickerson, R. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125,* 737–759. https://doi.org/10.1037/0033-2909.125.6.737

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., Hock, H., & Halberda, J. (2014). Hysteresis affects approximate number discrimination in young children. *Journal of Experimental Psychology: General, 143,* 255–265. https://doi.org/10.1037/a0030825

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'Grady, S., & Xu, F. (in press). The development of nonsymbolic probability judgments in children. *Child Development*. https://doi.org/10.1111/cdev.13222

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

Parsons, S. (2020). *splithalf: Robust estimates of split half reliability*. https://doi.org/10.6084/m9.figshare.11956746.v4

Paulus, M., Proust, J., & Sodian, B. (2013). Examining implicit metacognition in 3.5-year-old children: An eye-tracking and pupillometric study. *Frontiers in Psychology, 4,* 145. https://doi.org/10.3389/fpsyg.2013.00145

Paulus, M., Tsalas, N., Proust, J., & Sodian, B. (2014). Metacognitive monitoring of oneself and others: Developmental changes during childhood and adolescence. *Journal of Experimental Child Psychology*, *122*, 153–165. https://doi.org/10.1016/j.jecp.2013.12.011

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

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science, 306*, 499–503. https://doi.org/10.1126/science.1102085

Pierce, C. S., & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences, 3*, 77–83. https://doi.org/10.1093/mind/os-XI.41.128-a

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

Poulin-Dubois, D., & Brosseau-Liard, P. (2016). The developmental origins of selective social learning. *Current Directions in Psychological Science, 25*, 60–64. https://doi.org/10.1177/0963721415613962

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*, 515–526. https://doi.org/10.1017/S0140525X00076512

Proust, J. (2012). Metacognition and mindreading: One or two functions? In M. J. Beran, J. Brandl, J. Perner, & J. Proust (Eds.), *The Foundations of Metacognition* (pp. 234–251). Oxford University Press.

Pun, A., Birch, S. A. J., & Baron, A. S. (2016). Infants use relative numerical group size to infer social dominance. *Proceedings of the National Academy of Sciences, 113,* 2376–2381. https://doi.org/10.1073/pnas.1514879113

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J. M., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T., … Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4,* 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision-making. *Psychological Science*, *26,* 1664–1680. https://doi.org/10.1177/0956797615595037

Resendes, T., Elkaim, B., & Poulin-Dubois, D. (2019, October). *Do metacognitive strategies predict social selective learning in preschoolers?* [Poster Presentation]. Cognitive Development Society Biennial Meeting, Louisville, KY.

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS One*, *9,* e98663. https://doi.org/10.1371/journal.pone.0098663

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27,* 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Roebers, C. M., Kälin, S., & Aeschlimann, E. A. (2019). A comparison of non-verbal and verbal indicators of young children's metacognition. *Metacognition and Learning, 15,* 31–49. https://doi.org/10.1007/s11409-019-09217-4

Rohwer, M., Kloo, D., & Perner, J. (2012). Escape from metaignorance: How children develop an understanding of their own lack of knowledge. *Child Development, 83*, 1869–1883. https://doi.org/10.1111/j.1467-8624.2012.01830.x

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*, e17. https://doi.org/10.1017/pen.2018.16

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

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

Selmeczy, D., & Ghetti, S. (2019). Here is a hint! How children integrate reliable recommendations in their memory decisions. *Journal of Experimental Child Psychology*, *177*, 222–239. https://doi.org/10.1016/j.jecp.2018.08.004

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*, 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Shaw, A., Montinari, N., Piovesan, M., Olson, K. R., Gino, F., & Norton, M. I. (2014). Children develop a veil of fairness. *Journal of Experimental Psychology: General*, *143*, 363–375. https://doi.org/10.1037/a0031247

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, *141*, 382–395. https://doi.org/10.1037/a0025907

Shea, N., & Frith, C. D. (2019). The global workspace needs metacognition. *Trends in Cognitive Sciences, 23,* 560–571. https://doi.org/10.1016/j.tics.2019.04.007

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

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review, 120,* 779–797. https://doi.org/10.1037/a0034191

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition, 96,* B1-11. https://doi.org/10.1016/j.cognition.2004.07.004

Song, C., Kanai, R., Fleming, S., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition, 20,* 1787–1792. https://doi.org/10.1016/j.concog.2010.12.011

Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language. In *Language in mind: Advances in the study of language and thought* (pp. 277–311). MIT Press.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10,* 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science, 348,* 91–94. https://doi.org/10.1126/science.aaa3799

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition, 163,* 1–14. https://doi.org/10.1016/j.cognition.2017.02.008

Sumner, E., DeAngelis, E., Hyatt, M., Goodman, N., & Kidd, C. (2019). Cake or broccoli? Recency biases children's verbal responses. *PLoS One, 14,* e0217207. https://doi.org/10.1371/journal.pone.0217207

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*, 444. https://doi.org/10.3389/fpsyg.2013.00444

Taylor, M., Esbensen, B. M., & Bennett, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child Development, 65,* 1581–1604. https://doi.org/10.2307/1131282

Thompson, R. A., & Lagattuta, K. H. (2006). Feeling and understanding: Early emotional development. In *Blackwell Handbook of Early Childhood Development* (pp. 317–337). Blackwell Publishing. https://doi.org/10.1002/9780470757703.ch16

Tong, Y., Wang, F., & Danovitch, J. (2020). The role of epistemic and social characteristics in children's selective trust: Three meta-analyses. *Developmental Science, 23,* e12895. https://doi.org/10.1111/desc.12895

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances*, *2*, 2398212818810591. https://doi.org/10.1177/2398212818810591

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134,* 383–403. https://doi.org/10.1037/0033-2909.134.3.383

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). Taylor & Francis.

van Loon, M., de Bruin, A. B. H., van Gog, T., & van Merriënboer, J. J. G. (2013). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning*, *8*, 173–191. https://doi.org/10.1007/s11409-013-9100-0

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

van Loon, M., Destan, N., Spiess, M. A., de Bruin, A., & Roebers, C. M. (2017). Developmental progression in performance evaluations: Effects of children's cue-utilization and self-protection. *Learning and Instruction*, *51*, 47–60. https://doi.org/10.1016/j.learninstruc.2016.11.011

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582–600. https://doi.org/10.1037/0278-7393.26.3.582

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*, 159–176. https://doi.org/10.1016/j.lindif.2004.12.001

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14. https://doi.org/10.1007/s11409-006-6893-0

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

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wang, J. J., Odic, D., Halberda, J., & Feigenson, L. (2016). Changing the precision of preschoolers' approximate number system representations changes their symbolic math performance. *Journal of Experimental Child Psychology, 147*, 82–99. https://doi.org/10.1016/j.jecp.2016.03.002

Weber, E. H. (1978). *The sense of touch* (1st ed.). Academic Press for Experimental Psychology Society.

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind tasks. *Child Development, 75*, 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Winman, A., Juslin, P., Lindskog, M., Nilsson, H., & Kerimi, N. (2014). The role of ANS acuity and numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology*, *5*, 851. https://doi.org/10.3389/fpsyg.2014.00851

Wong, H., & Odic, D. (2020). The intuitive number sense contributes to symbolic equation error detection abilities. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000803

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*, 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

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

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272. https://doi.org/10.1037/0033-295X.114.2.245