Are Children's Judgments of Another's Accuracy Linked to Their Metacognitive Confidence Judgments?

Carolyn Baer, University of British Columbia, ORCID: 0000-0003-4521-8006

Puja Malik, University of British Columbia

Darko Odic, University of British Columbia, ORCID: 0000-0001-5783-2789

Contact Info:
Carolyn Baer
Department of Psychology
University of British Columbia
2136 West Mall
Vancouver, British Columbia V6T 1Z4
Email: cebaer@psych.ubc.ca
Word Count: 15648

Acknowledgements: This work was funded by the Social Sciences and Humanities Research Council of Canada through an Insight Development Grant to DO and a Canada Graduate Scholarship to CB. Thanks to Stephanie Lee, Samantha Kwan, Charlotte Aitken, and Eliscia Sinclair for their help with data collection, and to the schools and families for their support.

Abstract

The world can be a confusing place, which leads to a significant challenge: how do we figure out what is true? To accomplish this, children possess two relevant skills: reasoning about the likelihood of their own accuracy (metacognitive confidence) and reasoning about the likelihood of others' accuracy (mindreading). Guided by Signal Detection Theory and Simulation Theory, we examine whether these two self- and other-oriented skills are one in the same, relying on a single cognitive process. Specifically, Signal Detection Theory proposes that confidence in a decision is purely derived from the imprecision of that decision, predicting a tight correlation between decision accuracy and confidence. Simulation Theory further proposes that children attribute their own cognitive experience to others when reasoning socially. Together, these theories predict that children's self and other reasoning should be highly correlated and dependent on decision accuracy. In four studies (N = 374), children aged 4-7 completed a confidence reasoning task and selective social learning task each designed to eliminate confounding language and response biases, enabling us to isolate the unique correlation between self and other reasoning. However, in three of the four studies, we did not find that individual differences on the two tasks correlated, nor that decision accuracy explained performance. These findings suggest self and other reasoning are either independent in childhood, or the result of a single process that operates differently for self and others.

Keywords: Confidence, selective social learning, signal detection theory, simulation theory, individual differences

Declarations

Funding

This work was funded by the Social Sciences and Humanities Research Council of Canada through an Insight Development Grant to DO and a Canada Graduate Scholarship to CB.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Ethical Procedures

These studies were approved by the University of British Columbia Behavioural Research Ethics Board [H14-01984]. Parents/guardians of children in the study provided informed consent.

Availability of Data and Material

Stimuli, data, and associated extraction and analysis scripts (in R) are available on OSF.

Are Children's Judgments of Another's Accuracy Linked to Their Metacognitive Confidence Judgments?

How do children distinguish truths from falsehoods, like determining that their sibling is lying about the moon being made of cheese? As early as infancy, humans possess several tools to help evaluate truthfulness, including early-emerging core concepts for physical objects and psychological agency (Carey, 2009; Spelke & Kinzler, 2007), and a tendency to trust others (Csibra & Gergely, 2009; Harris, 2012). While these cognitive tools are generally helpful, over-reliance on any single one could lead to mistaken beliefs about the world. For example, if we relied only on our pre-existing concepts, we might never learn counterintuitive truths like the Earth's sphericity. If we relied only on trusting others' knowledge, we could fall prey to mischievous lies about cheesy moons. To avoid this, children and adults must consider the reliability of their own knowledge and that of the information provided by others to determine which ideas and concepts to trust and which to disregard.

Accordingly, children have at least two abilities that help them evaluate the reliability of evidence, one focused on the self and one focused on others. First, in part through their broader metacognitive toolbox, children can reason about their own confidence: the graded signal of whether an answer is likely to be true (Flavell, 1979; Pouget et al., 2016). For instance, our sense of confidence might tell us we are likely or unlikely to land a jump, that we might need to double-check an answer on an exam, or that we're pretty sure the moon is made of rocks. As some evidence for this ability in childhood, preschoolers report higher confidence when they correctly identify a pixelated object or remember seeing an object than when they are incorrect (Hembacher & Ghetti, 2014; K. E. Lyons & Ghetti, 2011) and can both prospectively and

retrospectively judge their accuracy on simple perceptual discrimination tasks (Baer & Odic, 2019). Twenty-month-old infants similarly respond to uncertainty by asking their caregiver for help (Goupil et al., 2016). These findings suggest that reasoning about confidence, evaluating the reliability of one's own knowledge, emerges at a young age.

Second, children also detect the reliability of knowledge in *others*. Decades of research in *mindreading* (also known as mentalizing or Theory of Mind) demonstrates that by at least age 3, children detect ignorance in others and use this to predict another's behavior or choose when to offer help (Liszkowski et al., 2008; Onishi & Baillargeon, 2005; Wimmer & Perner, 1983). Infants and children also make strategic decisions about which people to believe by tracking past accuracy, group membership, confidence displays, and other cues to accuracy (Birch et al., 2010; Koenig et al., 2004; Mills, 2013; Poulin-Dubois & Brosseau-Liard, 2016), which could help children discount statements about cheesy moons from lying siblings. For instance, when only one of two teachers consistently gives the correct labels for objects, preschool children selectively trust the accurate labeller's future answers and ask her for help instead of the inaccurate labeller (Koenig et al., 2004). Children are therefore also sensitive to the likelihood of others' knowledge when discerning the truth, rather than being entirely gullible to incoming information.

While extensive work has shown both that children have a metacognitive ability to evaluate their own confidence and are sophisticated in how they evaluate the reliability of others, these two research programs have remained separate, in part because they concern distinct targets: the reliability of self vs. another. However, several theorists have pointed out the potential for overlap between self and other reasoning (Carruthers, 2009; Gopnik, 1993; Proust, 2012). Most notably, reasoning about the self and about others involves thinking about an

individual's knowledge, and specifically the likelihood that the knowledge in question is correct. Accordingly, children could base both self and other evaluations on similar information (e.g., noting for themselves and also for others that long decision times signal answers that are more likely to be wrong; Koriat & Ackerman, 2010b), and could therefore rely on similar processing mechanisms when using this information to form reliability estimates in both cases.

In support, several studies have documented correlations and commonalities between self-focused metacognitive abilities (including reasoning about confidence specifically) and other-focused mindreading abilities (e.g., Gopnik & Astington, 1988; Kuzyk et al., 2020; Lecce et al., 2015; Lockl & Schneider, 2007; Paulus et al., 2014; but see Bernard et al., 2015; van Loon & van de Pol, 2019). In one study, three-year-old's mindreading abilities measured through a battery of false belief tasks predicted their metacognitive knowledge of memory strategies (e.g., spending more time to memorize difficult items) at age 5 (Lockl & Schneider, 2007). Another recent study presented 18-month-old infants with non-verbal self and other reasoning tasks, using a measure of persistence to index confidence (Kuzyk et al., 2020). Infants who poorly monitored their confidence (i.e., persisted on tasks that had no probability of success) were also more likely to learn a new object label from an unreliable teacher (a sign of poor social reasoning; Kuzyk et al., 2020). At the same time, however, some studies find no correlations between reasoning about the self and others. Bernard and colleagues (2015) found no correlation between preschool children's metacognition in an opt-out paradigm (where they can choose not to answer when uncertain) and their performance on false belief tasks, a finding replicated in a cross-cultural sample of German and Japanese children (Kim et al., 2020). It thus remains unclear when, if ever, metacognition and mindreading should correlate in childhood and what that means for the way the two abilities operate in the mind.

There are many possible explanations of how reasoning about the self and about others could be linked, and we devote some space in the General Discussion to those not tested here. Here, we test one relatively simple explanation: both judgments about one's own confidence and about another's likely accuracy are rooted in children's understanding of the task that the self and other are doing. Children who understand the task well are better positioned to notice deviations in knowledge and performance in both the self and others, while children who do not understand the task well are not. Importantly, this explanation predicts that there should be correlations in reasoning about expected accuracy when the self and other are doing the same task, but not necessarily when completing different tasks.

This potential mechanistic link between reasoning about the reliability of self and others that is rooted in Signal Detection Theory (SDT) and Simulation Theory. SDT, a highly influential theory in the fields of perception and memory, is designed to explain decision-making under uncertainty (Green & Swets, 1966). The critical observation of SDT is that our internal representations of the world are imprecise, leading to a degree of uncertainty in every decision we make (e.g., what we estimate to be 300 words on a page could plausibly be 237 or 413 or anything within that range; Dehaene, 2011; Green & Swets, 1966). When there is more variability in our decision (e.g., a range of 100-1000 words), we should be less confident about our accuracy; when there is more precision in that decision (e.g., a much narrower range of 299-301 words), we should experience much more confidence. In fact, SDT proposes that our confidence should be *directly* proportional to the degree of variability (Alais & Burr, 2004; Mamassian, 2016). In this way, confidence is much like a standard deviation around a mean, which quantifies the variability or uncertainty around the true mean (fittingly leading to a larger "confidence interval"): when the information we get is imprecise, we should lower our

confidence so as not to trust incorrect information. Because SDT provides a computationally simple mechanism for reasoning about confidence, it is an appealing account of how metacognitive confidence evaluations arise (Galvin et al., 2003; Kiani & Shadlen, 2009; Mamassian, 2016; Maniscalco & Lau, 2012; Pleskac & Busemeyer, 2010). Further, it can easily be tested by manipulating the amount of imprecision in a given decision, say by masking an item to be identified or by changing the ratio between two magnitudes to be close (high imprecision) or far (low imprecision).

On its own, SDT makes no explicit prediction about whether reasoning about confidence in the self vs. others is a single process or not; it simply states that if you have an imprecise internal representation, your expected accuracy (confidence) can be estimated from the amount of internal variability. However, a popular theory within the mindreading literature, *Simulation Theory* (Goldman, 2006; Jost et al., 1998; Meltzoff, 2007; Nickerson, 1999), could accommodate this link. Under this view, interpreting others' mental states occurs through a simulation process, metaphorically putting oneself in another's shoes, and attributing the experienced mental states to that other. When a child is reasoning about another's likelihood of accuracy, then, this simulation would involve reasoning metacognitively about their *own* likelihood of accuracy but attributing that likelihood to another person.

As evidence for Simulation Theory, first-hand experience seems to enable children to detect mental states in others when they otherwise wouldn't, suggesting that reasoning about the self is intricately linked to and predictive of reasoning about others. For instance, 3-month-old infants who gain experience grasping a desired object by wearing sticky mittens (something otherwise difficult for 3-month-olds) can then detect the goals of a grasping hand (Sommerville et al., 2005), even though 3-month-olds without this experience do not infer these goals

(Woodward, 1998). 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 predicted 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 who have learned and forgotten trivia items are more likely to attribute knowledge of those items to their peers, but not knowledge of items they were never taught (Birch et al., 2017). These findings strongly suggest that our evaluations of others – attributing goals, judging others' learning, and estimating the prevalence of knowledge – are highly influenced by reasoning about the self.

In combination, then, SDT and Simulation Theory together could explain correlations between own and other reasoning by arguing that both abilities rely on the same key process of quantifying the imprecision of a decision. That imprecision can be attributed to the self as feelings of confidence or to another as an assessment of likely accuracy. For example, if we estimate that there are between 200 and 300 words on this page, we could – using principles of SDT – judge our own confidence of the decision that there are "1000 words" as very low in probability. Similarly, if we hear somebody else estimating that there are "1000 words" on this page, we could likewise use the same process to estimate our confidence in *their* decision, judging them to be an unreliable teacher for future word-estimation decisions. Therefore, any time a child and the person they are observing have access to the same information (e.g., are looking at the same page of words), we should expect to find strong correlations in their estimates of self and other accuracy. This is often true in the studies reporting correlations – children evaluate whether they know what an object really is and then evaluate another's

perspective about the same object (Gopnik & Astington, 1988), or they evaluate whether they or another know the answer to one target memory pair (Paulus et al., 2014).

Besides making identical judgements when given identical information, the combination of Simulation Theory and SDT also makes a second testable prediction. Since SDT proposes that confidence is a direct computation of the imprecision of a decision and Simulation Theory proposes that evaluating another's likely accuracy is the same as reasoning about your own, the combination of these accounts predicts that individual differences in the imprecision of decisions should be the core source of variability for both self and other judgments. Put differently, if we can directly measure a subject's decision imprecision and statistically control for it, any correlations between self and other reasoning should disappear.

Therefore, in the studies reported here, we set out to test these two predictions: (1) whether there are strong correlations between detecting accuracy in the self and others when given access to the same task (and not when using an unrelated task), and (2) whether these correlations are eliminated when controlling for individual differences in the imprecision of decisions in that task. We chose to use a task tapping into children's perception of area, an early-developing ability used previously in metacognitive tasks (Baer et al., 2018; Baer & Odic, 2020a; Salles et al., 2016). This task allows us to experimentally manipulate the imprecision of the decision by making the shapes close in size (harder, more imprecision) or disparate (easier, less imprecision), something much harder to achieve in the memory tasks used in past work. We correlated how well children computed the imprecision in simple area discrimination decisions

(i.e., how sure am I that I know which of two shapes is larger) compared to their judgements of how well *others* do it (i.e., which of two agents did better on a shape drawing competition).[1]

In designing this study, we also needed to ensure that any correlations found are not the result of other processes in common to the self and other judgments. For example, many mindreading and metacognition tasks depend on the use of mental state verbs (e.g., "know" or "think"). While the shared language signals that judgments of knowledge in both the self and others contribute to a shared concept of 'knowledge', relying on this language may artificially link the processes leading to such judgments, which may themselves be distinct. As one solution, some recent studies have turned to measures of *procedural metacognition* (or reasoning about confidence without requiring mentalistic language). For example, the study by Kuzyk and colleagues (2020) described earlier uses a measure of persistence rather than explicit report, thereby avoiding the confound of developing mental state language that was present in past work. Therefore, following this work, we also chose paradigms that measure children's self and other reasoning in a way that avoids mentalistic language.

A second major problem is the influence of response biases. Within the metacognition literature, it is well documented that children tend to report higher confidence in their knowledge and abilities than is warranted (e.g., Destan & Roebers, 2015; Hagá & Olson, 2017; Taylor et al., 1994; van Loon et al., 2017), possibly due to overoptimistic beliefs about the self (Lockhart et al., 2017). Similarly, when reasoning about others' knowledge, children will imitate seemingly irrelevant actions even though they seem to understand the irrelevance (D. E. Lyons et al., 2007; Meltzoff, 1988) and trust adults who blatantly lie to them even when children know the truth

[1] Note that because our main task uses a 2-alternative forced choice format (choosing which shape is larger), we are specifically referring to theories of SDT that propose confidence is a relative computation of decision imprecision. Other variations of SDT theories are briefly discussed in the General Discussion.

(Jaswal, 2010). Given these patterns, which co-occur in childhood (Hagá & Olson, 2017), there is a chance that some of the reported correlations between self and other evaluations might stem from common response biases. For instance, in Kuzyk et al. (2020), some infants might be naturally inclined to seek out as much information as possible when faced with uncertainty both by persisting 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. Other response biases, like a desire to please adults or to be optimistic, could similarly affect both self and other judgments and lead to correlations even if the two have completely unrelated cognitive processes.

In the current work, we test for correlations between self and other reasoning about accuracy using tasks that eliminate overconfident response biases and mental state language while maintaining the critical commonality deemed necessary under the SDT/Simulation Theory account: task-specific decision imprecision. First, we adopt a relative choice paradigm for both the self and other tasks, rather than relying on absolute judgments like declarations of knowledge or decisions to trust or not trust a teacher. In a relative task, children are asked to indicate which of two options best fits a given criteria (in this case, which is more likely to be true). Notice that in doing so, we remove a child's ability to respond overconfidently because they must pick one of two answers they cannot simply say 'yes' to everything. Instead, they must reason about which of the two options is *the most* likely to be true. In the self reasoning task, this involves selecting which of two questions children feel most sure of answering correctly (e.g., Baer & Odic, 2019; Butterfield et al., 1988). In the other reasoning task, this involves selecting which of two teachers children feel is more reliable (called a selective social learning task; e.g., Birch et al., 2008; Einav & Robinson, 2010; Koenig et al., 2004). By experimentally eliminating these

response biases, we can ensure that any correlations between the self and other judgments do not stem from shared response biases.

Then, to limit the use of mental state language, we ask children to make strategic judgments that rely on assessments of knowledge (see Crivello et al., 2018; Hembacher & Ghetti, 2014). For example, instead of asking children to report which teacher knows more, we ask children to make a strategic choice to ask one teacher for help (Einav & Robinson, 2010; Koenig et al., 2004). And, instead of asking children to report whether they know an answer or not, we ask them to strategically answer the question they feel most sure about (Baer & Odic, 2019). These changes help reduce the potential influence of shared linguistic concepts inducing correlations between the two tasks.

The prediction of the SDT/Simulation account is that self and other reasoning on these two tasks should correlate despite the removal of these third variable explanations, provided one critical condition is true. The SDT account requires that they will correlate if and only if they are both computed from the *same* decision imprecision, which is thought to be dimension-specific (Baer et al., 2018; Vo et al., 2014). That is, children's confidence in their own or another's estimate of the number of words on a page should be related, but both should be entirely unrelated to their confidence in their own or another's estimate of the emotional expression on a face. We therefore adapted both paradigms to use a single target decision. Specifically, children were asked to reason about the relative sizes of shapes, tapping into a system of representing area that is well-developed in childhood but still subject to individual differences (Brannon et al., 2006; Odic, 2018). In fact, these individual differences in area representation are thought to be the direct result of representational imprecision – the more imprecise a child's perceptual representation of area, the harder it will be for that child to tell apart two sizes (Brannon et al.,

2006; Odic, 2018; Odic et al., 2013). Therefore, to test the second prediction that any correlation between the self and other tasks should be eliminated when controlling for imprecision, we can use children's accuracy on these area discriminations to capture the imprecision in their sense of area.

Together, by using two paradigms that eliminate response biases and mental state language as potential third variables, and by using a single type of representation in both paradigms, we can test whether SDT and Simulation Theory together can explain how children reason about the likely accuracy of their own and other's knowledge. In four studies, we test whether self and other reasoning is correlated when representations are shared, and uncorrelated when representations are distinct (the first prediction). Then, in Experiments 3 and 4, we use two techniques to test whether shared representations entirely explain any correlations (the second prediction). To anticipate our findings, we surprisingly failed to detect a consistent correlation between self and other reasoning in the studies – a basic requirement of the SDT/Simulation account. We therefore also conducted a mega-analysis of all four studies to clarify the results.

Experiment 1

Methods

Participants. A total of 80 children participated in the study (44 girls), meeting the planned sample size of 80 children (20 per age group, set arbitrarily a priori, see Simmons et al., 2011). We focused on children between 4 and 8 years (M = 5;11 [years; months], range = 4;0 - 7;10) to overlap the age ranges of studies using similar paradigms (Baer & Odic, 2019; Einav & Robinson, 2010). Both these studies show development in these skills over this age range, which additionally helps us find individual differences that should correlate between the two tasks

according to the SDT/Simulation theory account. Five additional children were tested but not included in the sample because they did not complete the study. 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. To test whether children's evaluations of self and other accuracy correlate when using the same representational variability, we adapted two existing paradigms: the selective social learning paradigm (reasoning about others) and the relative confidence task (reasoning about the self). Children completed the two tasks in a fixed order, with the selective social learning task first and the confidence discrimination task second. Both tasks were presented on a laptop.

Selective Social Learning Task. To assess individual differences in children's reasoning about others' knowledge, we used a selective social learning paradigm modified to rely on area representations. In the selective social learning paradigm, two informants are shown to differ on a critical trait (e.g., past accuracy, displayed confidence, social group, etc.) and children are asked to make a series of social judgments to indicate which informant they find most reliable. If children consistently choose one informant over the other, we can reason that children notice and care about the critical trait (e.g., Birch et al., 2010; Koenig et al., 2004). Here, we based our task on a variation by Einav and Robinson (2010) in which one informant is consistently more accurate than the other during a short 'history' phase. Importantly, neither informant is truly accurate – both informants provide incorrect answers that differ only in the magnitude of their error. For example, in Einav and Robinson's study, informants said that either 6 or 10 dots were on a card when there were actually 5. Because children cannot evaluate accuracy here by

identifying the correct or incorrect item (as both are incorrect), children must rely on their evaluations of the magnitude of each informant's error, i.e., the relative likelihood of their accuracy.

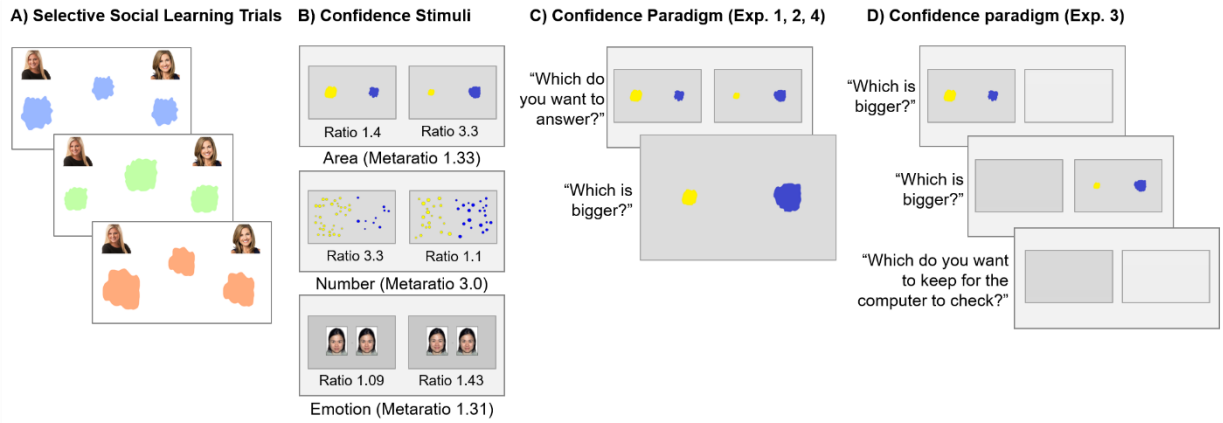
In our modified version using area representations, children saw photos of a pair of "contestants in a drawing contest" and were asked to help the experimenter "choose the winner". At the beginning of the study, the experimenter told children that the contestants had to copy a shape perfectly, and that 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 (Le Corre & Carey, 2007).

To obtain coarse individual differences, children completed 4 trials, each with three judgments of informant accuracy. In each trial, children were introduced to a new pair of White female "contestants". Because the SDT/Simulation account proposes that only representational imprecision is used to compute confidence, we used only photographs of the contestants rather than videos or live actors to remove other potential cues to confidence like reaction time or movement cues (see Kominsky et al., 2016 for a similar method). Children then saw three drawings from each contestant which critically differed only in size. Each example started with a target shape in its true size in the center of the screen, followed by the "copies" made by each contestant underneath their respective photos (see Figure 1). Across the three examples within each trial, one contestant consistently produced shapes that were relatively closer in size to the target (either larger or smaller than the target by a ratio of 1.2; e.g., 120% or 83% of the original size), while the other contestant produced relatively further-sized shapes (a ratio of 2.0 – 200% or 50% – over or under in the same direction as the 'closer' contestant). 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 left/right positioning of the closer-sized shape was counterbalanced across the 4 trials, and the identities of the 'winning' contestants were counterbalanced between participants.

Figure 1

Selective Social Learning Task and Confidence Task Stimuli.



Note. Panel A depicts a sample trial in the Selective Social Learning Task. Children saw three examples of two informants' drawings beside a target object. The right contestant is the 'closer' contestant in this example. Children then made Winner, Ask, and Endorse judgments about this pair of contestants. Panel B depicts sample stimuli from the Confidence Task in Experiment 1. Only the Area trials were used in Experiments 2-4. Panel C depicts the confidence paradigm used in Experiments 1, 2, and 4. Children first selected which of two questions they wanted to answer, then answered only that question. Panel D depicts the confidence paradigm used in Experiment 3. Children first answered each question, then selected which answer they felt most confident about.

Following the three examples in each trial, children answered three test questions based on classic selective social learning measures (e.g., Birch et al., 2008; Einav & Robinson, 2010; Koenig et al., 2004). 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 relative accuracy. Children were then told they could

ask one of the two contestants for assistance on a drawing contest to be held in class, 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 the same size 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 Relative Confidence Task from Baer, Gill, and Odic (2018). In each trial, children had to make a simple choice, like whether a yellow or blue shape is larger (see Figure 1 and Odic, 2018). Critically, and following the principles of SDT, this task manipulates the degree of confidence participants should feel in the choices by varying specific properties of the stimuli. For instance, in the case of area judgments, when the ratio of pixels in the blue and yellow shapes is large like ratio 3.3 (e.g., 119,130 yellow pixels and 36,100 blue pixels), participants should experience higher confidence than in smaller ratios like 1.05 (Baer et al., 2018).

To assess sensitivity to these differences in confidence, children saw screenshots of area comparisons in pairs on the screen prior to answering and selected which of the two screenshots they "wanted to answer," appealing to a desire to answer correctly. Effectively, then, we were asking children to compare the sizes of shapes in two questions, then compare their confidence in each of those size comparisons. Children then answered only the selected question. Screenshots were 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). 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.

The confidence task included 3 independent perceptual dimensions: area, number, and emotion (detailed below; Baer et al., 2018; Odic, 2018; Vo et al., 2014). By extracting confidence judgments from children on each of the three dimensions, we can test the first prediction of the SDT/Simulation account that self and other reasoning should *only* correlate when representations are shared. That is, we would expect area confidence reasoning to correlate with our area-focused Selective Social Learning Task, but not number confidence or emotion confidence[2].

On each Area question, children selected whether a yellow or blue shape was bigger (see Figure 1 and Odic, 2018). Expected confidence in these choices was manipulated through the ratio of pixels in the blue and yellow shapes (e.g., a ratio of 3.3 yellow pixels for every blue pixel for high confidence or a ratio of 1.05 for low confidence; Baer et al., 2018). There were 5 ratios in total for this task (3.3, 2.1, 1.4, 1.1, and 1.05), which were then paired into three metaratios (3.0, 2.0, and 1.33). Metaratio pairs were created such that the largest shape was not always in the high confidence pairing to prevent children from using a heuristic like 'choose the largest of the 4 visible shapes.' On each Number question, children selected whether a set of yellow or blue

[2] Baer et al. (2018) did find correlations between these three tasks in children aged 6-9, slightly older than we tested here. However, other findings using these same dimensions (e.g., Vo et al., 2014) and using other dimensions (Bellon et al., 2020; Geurten et al., 2018) show no correlations between metacognitive dimensions under age 8.

dots was more numerous (see Figure 1 and Halberda et al., 2008). Here, expected confidence was manipulated through the ratio of dots in the blue and yellow sets using the same ratios as the area questions (3.3, e.g., 33 yellow dots and 10 blue dots, 2.1, 1.4, 1.1, and 1.05). Metaratio pairs, like the area trials, were 3.0, 2.0, and 1.33. In each Emotion question, children selected which of two expressions was happier (see Figure 1 and Baer et al., 2018; Vo et al., 2014). The expressions, taken directly from Baer et al. (2018), were created by blending a happy and angry expression by one of four female models (two Caucasian and two East Asian). The blended expressions ranged from 100% happy (i.e., 0% angry), through 53.3% happy (i.e., 46.7% angry). Expected confidence was manipulated by varying the ratio of the happy/angry weights (e.g., 93.3% happy vs. 60% happy, a ratio of 1.56 for a high confidence trial, 73.3% happy vs. 66.7% happy, a ratio of 1.1 for a low confidence trial). This resulted in 5 different binned ratios (1.09, 1.2, 1.31, 1.43, and 1.57), paired to make metaratios of 1.44, 1.31, and 1.1. See https://osf.io/dtzpq/?view_only=8c28abc81c7b47f6a234df1c36dbc5f0 for exact stimuli.

In past work with this paradigm, 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 (Baer et al., 2018; Baer & Odic, 2019). 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 than expected, the underlying ability to detect differences in confidence that we are interested in (i.e., to identify which trial feels ‘easy’ and which feels ‘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 that expects children to be increasingly

likely to select the easier question as the difference in difficulty increases. If children become *less* likely to select the easier question (i.e., they consistently pick the harder question), their data will only fit an inverted model, allowing us to identify these children and invert their data (see Baer et al., 2018 for a description of the model).

The Confidence Task consisted of 45 trials in total (five at each of three metaratios in the three dimensions), with 4 warm-up trials of the Area, Number and Emotion tasks alone (i.e., with no preceding confidence choice). To keep children engaged, the three dimensions were randomly intermixed and children received pre-recorded feedback from the computer ("Yeah, that's right!", "Oh, that's not right.") when they answered the Area, Number, and Emotion questions (e.g., which shape was larger). They received no feedback about their confidence choices (see Baer & Odic, 2019 for evidence that this feedback does not affect performance relative to neutral affirmations).

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 6 and 7, children reliably chose the closer contestant as the Winner, though children at age 4 did not (see Table 1 for means and tests against chance of 50%). Performance correlated with age, $r(78) = .48$, $p = .001$, indicating that older children were better at detecting and attributing the differences in sizes to the contestants. Similarly, 7-year-olds reliably Asked the closer contestant for help, while children aged 4 through 6 did not (see Table 1). There was a significant correlation with age, $r(78) = .24$, $p = .036$, suggesting that older children were more likely to use their judgments of error magnitude to inform their help-seeking

Table 1

Tests Against Chance at Each Age Group in Experiments 1 and 2.

	Experiment 1					Experiment 2				
	Mean (%)	SD	*t*(79)	*p*	*d*	Mean (%)	SD	*t*(80)	*p*	*d*
Winner										
4	56.25	21.27	1.31	.204	0.29	40.00	26.16	-1.71	.104	0.38
5	57.14	23.90	1.37	.186	0.30	64.29	26.89	2.43	.024	0.53
6	77.63	24.85	4.85	< .001	1.11	86.25	28.65	5.66	< .001	1.27
7	85.00	23.51	6.66	< .001	1.49	86.25	23.61	6.87	< .001	1.54
Ask										
4	51.25	23.61	0.24	.815	0.05	47.50	22.80	-0.49	.629	0.11
5	59.52	26.78	1.63	.119	0.36	67.86	23.90	3.42	.003	0.75
6	60.53	34.68	1.32	.202	0.30	63.75	33.91	1.81	.086	0.41
7	72.50	27.98	3.60	.002	0.80	80.00	28.79	4.66	< .001	1.04
Endorse										
4	50.00	25.65	0.00	1.00	0.00	41.25	24.70	-1.58	.130	0.35
5	57.14	27.55	1.19	.249	0.26	46.43	31.90	-0.51	.614	0.11
6	52.63	18.44	0.62	.542	0.14	52.50	35.26	0.32	.755	0.07
7	62.50	34.89	1.60	.126	0.36	48.75	32.92	-0.17	.867	0.04
Area Confidence										
4	63.33	14.43	4.13	.001	0.92	65.00	12.40	5.41	< .001	1.21
5	67.94	8.85	9.29	< .001	2.03	68.57	15.94	5.34	< .001	1.17
6	69.12	14.09	5.92	< .001	1.36	66.67	14.18	5.26	< .001	1.18
7	77.00	13.59	8.89	< .001	1.99	69.33	14.73	5.87	< .001	1.31
Number Confidence										
4	61.33	10.28	4.93	< .001	1.10					
5	62.22	11.42	4.91	< .001	1.07					
6	65.26	14.33	4.64	< .001	1.07					
7	67.33	12.96	5.98	< .001	1.34					
Emotion Confidence										
4	60.00	11.03	4.05	.001	0.91					
5	57.78	11.61	3.07	.006	0.67					
6	62.81	13.21	4.23	.001	0.97					
7	61.00	11.70	4.20	< .001	0.94					

decisions. However, we found a 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, $r(78) = .18$, $p = .114$. This was unexpected given

that a recent meta-analysis conducted on selective social learning tasks found strong evidence that children endorse informants they believe to be more accurate (Tong et al., 2020).[3] See the Supplemental Material for additional correlations between measures. Given this result, we report all correlations between this task and our confidence measure separately for each question.

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 in all three dimensions (see Table 1 for means and tests against chance). For these analyses, children who responded consistently with the harder option (4 in the Area trials: one 4-year-old, one 6-year-old, two 7-year-olds, 18 in the Number trials: seven 4's, five 5's, two 6's, and four 7's, and 17 in the Emotion trials: five 4's, seven 5's, one 6, and four 7's, 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. See the Supplemental Material for additional analyses of this task.

Correlations of Individual Differences. Given that both tasks demonstrated as expected that children were sensitive to relative differences in their own (Confidence Task) or another's relative accuracy (Selective Social Learning Task, except the Endorse trials), we next looked for correlations of individual differences between the two tasks. As shown in Table 2, we surprisingly found no correlations between the Winner, Ask, or Endorse choices and any of the three dimensions for the confidence task when controlling for age.

[3] Note that these data and those from Experiment 2 are included as unpublished data in this meta-analysis.

Table 2

Correlations Controlling for Age in Experiments 1 and 2.

	Area Conf.	Number Conf.	Emotion Conf.
Experiment 1			
Winner	-.08	.00	-.05
Ask	-.05	-.14	.06
Endorse	.02	-.16	.05
Experiment 2			
Winner	.29*		
Ask	.36**		
Endorse	-.02		

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

Discussion

Contrary to our predictions, we found no correlations between the self and other tasks. Although this was expected for the individual differences in emotion and number confidence (as the Selective Social Learning task only tested area and the SDT account predicts that correlations should only occur when the task is the same for both self and other), it was contrary to our expectation for the area Confidence task.

Nevertheless, we replicated two findings in the existing literature on self and other evaluations, signalling that our self and other tasks worked as intended. First, we replicated the key findings of Einav & Robinson's (2010) magnitude of error selective social learning task, in which children relied on informants who provided relatively more accurate answers. This occurred despite three changes to the paradigm (the use of photographs instead of videos, judgments about area rather than numerosity, and without the use of a number line to track estimates), but did not replicate in 4-year-olds, in children under age 7 in the Ask trials, or in the Endorse trials. We also found that children as young as 4 years could reason about their relative

states of confidence for Area, Number, and Emotion judgments, younger than in past reports (Baer et al., 2018; Baer & Odic, 2019), though we did not see an influence of the difference in trial difficulties or significant improvements with age.

One possible explanation for the failed replication is that the length of the confidence task with all three dimensions fatigued children, making it difficult to capture reliable individual variability. Therefore, in Experiment 2, we conducted a replication with only the area trials of the confidence task to reduce possible fatigue effects.

Experiment 2

Methods

Participants. Eighty-one children participated in the study (M = 5;11, range = 4;0 - 7;10, 50 girls) in the same manner and geographical location as Experiment 1. None of the children had participated in Experiment 1.

Materials and Procedures.

Selective Social Learning Task. We used the same 'drawing contest' task from Experiment 1, but with one small change: we made the two shapes in the Endorse trials different sizes to make the contrasting answers more noticeable.

Confidence Task. We used only the 15 Area trials from the confidence task in Experiment 1 to reduce fatigue and potential task-switching effects.

Results

Selective Social Learning Task. Children aged 5-7 reliably chose the closer contestant as the Winner, though children at age 4 did not (see Table 1 for means and tests against chance of 50%), and choice correlated with age, $r(79) = .56$, $p < .001$. Similarly, children aged 5 and 7

reliably Asked the closer contestant for help, while children aged 4 and 6 did not (see Table 1). There was still a significant correlation with age, $r(79) = .38$, $p = .001$. Once again, no age group Endorsed the closer contestant above chance rates (see Table 1), and there was no correlation with age, $r(79) = .16$, $p = .153$. Additional correlations are reported in the Supplemental Material.

Confidence Task. Children at all age groups selected the easier trial more than 50% of the time (see Table 1 for means and tests against chance). Eleven children consistently selected the harder task (one 4-year-old, three 5-year-olds, three 6-year-olds, and four 7-year-olds), and their data has been inverted. Additional analyses are reported in the Supplemental Material.

Correlations of Individual Differences. In contrast to Experiment 1, and as predicted by the SDT/Simulation account, the Confidence Task correlated with both the Winner and Ask choices, but not with Endorse choices (see Table 2, all correlations controlling for age).

Discussion

These results provide preliminary evidence for the SDT/Simulation account, with a correlation between the self and other tasks that held when controlling for age. The pattern of findings in the confidence task was largely the same, potentially suggesting that fatigue alone is insufficient to explain the lack of correlation in Experiment 1. However, 5-year-olds in this study chose to ask the closer informant for help, whereas only 7-year-olds did so in Experiment 1. It may therefore be possible that children in this sample, for whatever reason, encoded the difference in accuracy more deeply.

At the same time, these correlations alone only support the first of the two predictions of the SDT/Simulation account. The assumption of the SDT/Simulation account is that this

correlation is driven by a common system of representing confidence based on variability in area representations (in this case). While we experimentally eliminated common response biases and the necessity of common mentalistic language as potential third variables, this study does not yet provide evidence that the common area representations are the sole connector between self and other judgments. To examine this directly in Experiment 3, we collected a measure of area discrimination performance as part of the Confidence Task by relying on a retrospective confidence judgement: children first answered two area discrimination decisions (i.e., which of two shapes is larger), and then afterwards decided which of the preceding two trials they were more confident on. By having a measure of both their area perception *accuracy* and their metacognitive *precision*, we can test the prediction that self and other reliability judgements are no longer correlated when controlling for area accuracy.

Experiment 3

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 Experiments 1 and 2. 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 the previous experiments. In addition to the two tasks described, 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). We had hoped to use this measure as a coarse approximation for general intelligence, but we had low rates of completion (26 of 81 participants). These data are reported in the Supplemental Material.

Materials and Procedures.

Selective Social Learning Task. We used the same 'drawing contest' task, 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 before, 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 modified the Confidence Task so that rather than making prospective judgments about their success, children evaluated their confidence retrospectively (Baer & Odic, 2019, 2020a). Like before, children saw four warm-up trials of the 'blobs game' (area discriminations) 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. Children did not receive feedback in any part of the task, as feedback immediately after their area answers (but before their confidence choice) would have eliminated the need for them to reason metacognitively.

Because all children had to answer the same 40 Area questions, we used their accuracy on these questions as a measure of the imprecision in their area representations. As outlined in the Introduction, the variability that is thought to index confidence under SDT models is the very same variability that makes it difficult to compare two representations together. If a child possesses a very precise sense of area, they might only have difficulty comparing very similar areas to one another. However, if a child possesses an imprecise sense of area, they will be unable to compare even areas that are dissimilar and easy for other children. Because our Area questions all involve comparing two areas, we can thus infer that children whose accuracy is low on the Area questions possess less precise representations than children with high accuracy.

Results

Selective Social Learning Task. Children aged 5 and 7 chose the informant with lesser errors as the winner more often than chance of 50% (see Table 3 for means and tests against chance), and choice correlated with age, $r(79) = .35$, $p = .001$. Six-year-olds similarly chose the informant with lesser errors, though did not reach traditional levels of significance (see Table 3). As in Experiment 1, only 7-year-old children chose to Ask this contestant for help significantly more than chance (see Table 3), behavior which also correlated with age, $r(79) = .22$, $p = .048$. Once again, no age group Endorsed the closer contestant above chance rates (see Table 3), and there was no correlation with age, $r(79) = .07$, $p = .528$. Additional correlations are reported in the Supplemental Material. Therefore, despite attempts to make the task easier for younger children, 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.

Table 3

Means and Tests Against Chance for the Selective Social Learning Task, Confidence Task, and Area Task in Experiments 3 and 4

	Experiment 3					Experiment 4				
	Mean (%)	SD	*t*	*p*	*d*	Mean (%)	SD	*t*	*p*	*d*
Winner										
4	51.19	23.02	0.24	.815	0.05					
5	66.25	24.70	2.94	.008	0.66	62.22	29.01	2.83	.007	0.42
6	63.75	30.86	1.99	.061	0.45	79.55	26.56	7.38	< .001	1.11
7	78.75	20.32	6.33	< .001	1.41	94.19	14.26	20.31	< .001	3.10
Ask										
4	54.76	24.52	0.89	.384	0.19					
5	58.75	27.24	1.44	.167	0.32	54.44	29.33	1.02	.315	0.15
6	60.00	32.85	1.36	.189	0.30	77.27	25.75	7.02	< .001	1.06
7	70.00	25.13	3.56	.002	0.80	88.95	18.34	13.93	< .001	2.12
Endorse										
4	47.62	24.88	-0.44	.666	0.10					
5	47.50	25.52	-0.44	.666	0.10	50.00	26.11	0.00	1.00	0.00
6	55.00	27.63	0.81	.428	0.18	64.77	31.12	3.15	.003	0.47
7	51.25	28.65	0.20	.847	0.04	66.86	28.72	3.85	< .001	0.59
Confidence										
4	59.76	8.29	5.40	< .001	1.18					
5	63.50	12.99	4.65	< .001	1.04	68.89	14.00	9.05	< .001	1.35
6	78.50	13.19	9.66	< .001	2.16	71.97	13.83	10.53	< .001	1.59
7	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	86.70	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

Area Task. Children at all ages accurately chose the larger shape well above chance levels (see Table 3), and accuracy increased with age, $r(79) = .33$, $p < .001$, replicating previous work (Odic, 2018).

Confidence Task. Children at all ages chose the easier trial as their most certain more often than expected by chance (see Table 3), with a significant correlation with area discrimination, $r(79) = .28$, $p = .011$. Five children consistently chose the harder trial (three 4-year-olds and two 5-year-olds) and their data has been inverted. Additional analyses to replicate documented effects in this task are reported in the Supplemental Material.

Correlations of Individual Differences. Replicating Experiment 1 but *not* Experiment 2, children's performance on the Confidence Task did not predict their Winner, $r(78) = .14$, $p = .227$, Ask, $r(78) = .08$, $p = .499$, or Endorse answers, $r(78) = .12$, $p = .270$, when controlling for age. Despite not seeing the predicted correlation, we felt that it could still be informative to examine how shared decision imprecision impacted performance. To do this, we conducted a hierarchical regression in two steps. In the first step, we included age and Area accuracy (our index of sensitivity to decision imprecision), and then added Confidence Task performance in the second step. The SDT/Simulation account predicts that Area accuracy should predict children's social judgments in the first step, and that the Confidence measure should add no additional variability in the second step, given that the area and confidence measures are thought to tap the same decision imprecision. As shown in Table 4, however, we did not find evidence supporting this prediction. In step 1, there was no meaningful contribution of area accuracy, with age serving as the only significant predictor. In step 2, there was no meaningful contribution of confidence performance over and above age and area accuracy, but because area accuracy was not a meaningful predictor, this is further inconsistent with the SDT/Simulation account prediction. It appears, then, that an age-related change is responsible for the correlations between tasks rather than shared area representations.

Table 4

Hierarchical Regressions in Experiments 3 and 4.

	Experiment 3								Experiment 4							
	Model Fit				Coefficients				Model Fit				Coefficients			
Model	R^2	ΔR^2	F	p	Predictor	ß	t	p	R^2	ΔR^2	F	p	Predictor	ß	t	p
DV: Winner																
Step 1	.12	.12	5.51	.006	Age	.35	3.10	.003	.19	.19	15.14	< .001	Age	.44	5.50	< .001
					Area Acc.	.01	0.10	.923					Area Acc.	-.07	-0.87	.386
Step 2	.14	.02	1.46	.231	Age	.26	1.99	.050	.19	.00	0.03	.864	Age	.44	5.42	< .001
					Area Acc.	-.01	-0.05	.963					Area Acc.	-.07	0.25	.385
					Confidence	.16	1.01	.231					Confidence	-.01	-0.17	.864
DV: Ask																
Step 1	.05	.05	2.08	.132	Age	.20	1.71	.091	.20	.20	16.08	< .001	Age	.45	5.66	< .001
					Area Acc.	.06	0.47	.637					Area Acc.	-.03	-0.35	.731
Step 2	.05	.00	0.40	.534	Age	.15	1.11	.269	.20	.00	.01	.920	Age	.45	5.57	< .001
					Area Acc.	.05	0.40	.694					Area Acc.	-.02	-0.35	.730
					Confidence	.09	0.62	.534					Confidence	-.01	-0.10	.919
DV: Endorse																
Step 1	.01	.01	0.51	.602	Age	.10	0.86	.393	.03	.03	2.00	.139	Age	.16	1.86	.065
					Area Acc.	-.09	-0.79	.433					Area Acc.	.05	0.53	.599
Step 2	.03	.02	1.47	.230	Age	.01	0.09	.931					Age	.16	1.76	.081
					Area Acc.	-.11	-0.93	.356					Area Acc.	.05	0.54	.593
					Confidence	.17	1.21	.230					Confidence	.03	0.38	.704

Discussion

While we replicated some results of Experiments 1 and 2 in that children were selective in who to trust, and were sensitive to differences in confidence, we saw some major differences that cast doubt on the SDT/Simulation account. As in Experiment 1, there was no correlation between self and other judgments, and a hierarchical regression revealed that there was no influence of area accuracy on the three Selective Social Learning Task measures.

Once again, only 7-year-olds, the oldest children in our sample, showed above-chance selectivity to ask the closer informant for help though even 5-year-olds choose the closer informant as the winner. Their answers were above chance, though not significantly so, potentially meaning that we lacked sufficient power to detect these effects. While this alone

cannot explain the pattern of correlations in the current experiment, we felt that it warranted a final study with a larger sample of children to provide a more definitive picture. A second possibility for the lack of correlations between self and other reasoning in Experiment 3 is the choice of a retrospective confidence task (as opposed to a prospective one in Experiments 1 and 2). Some have suggested that confidence signals derived before making a decision are distinct from those we make *after* a decision (Pouget et al., 2016), with confidence occurring before a decision relying more heavily on representational imprecision. To test both of these possibilities, we replicated Experiment 2, but assessed area accuracy through an additional task rather than embedding it into the confidence task as in Experiment 3.

Experiment 4

Methods

Participants. Using the observed correlation in Experiment 2 between children's choice of Winner and performance on the Confidence task, we calculated that 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 the other experiments. Three additional children were excluded for not completing the study in full, and none of the children in the sample had participated in the previous studies. As in Experiment 3, we asked parents to fill out the DVAP online within two weeks of participation. Forty-seven parents completed the assessment, and this data is reported in the Supplemental Material.

Materials and Procedures.

Selective Social Learning Task. We used the same stimuli as in Experiment 3. Because our previous studies had not found that children at any age were selecting the closer informant on Endorse questions, we decided to modify the wording of this question slightly to clarify what we were asking. We asked children "Which girl's shape would you guess looks the way it is supposed to look?," which signaled that it was permissible to indicate the same girl as in previous responses (since it is 'just a guess'), and highlighted the relation between the girls and the shapes in case children were only focusing on the features of the shapes.

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 with only Area trials from Experiment 2, but without the warm-up trials as all children completed the area task first.

Results

Selective Social Learning Task. As shown in Table 3, 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 the previous studies also Endorsed her shape as being closer, while 5-year-olds did neither (see Table 3). Ask choices correlated with age, $r(130) = .45$,

$p < .001$, and Endorse choices had trending correlation with age, $r(130) = .17$, $p = .055$. Additional correlations are reported in the Supplemental Material.

Area Task. Children in all three age groups successfully identified the larger shape well above chance levels (see Table 3), and accuracy did not significantly increase with age, $r(130) = .11$, $p = .213$.

Confidence Task. Children at all ages once again chose the easier question more than expected by chance (see Table 3). Eleven children consistently chose the harder trial (five 5-year-olds, two 6-year-olds, and four 7-year-olds) and their data has been inverted. There was no correlation with area performance, $r(130) = -.01$, $p = .939$.

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, $\beta_{Age} = .18$, $t(129) = 2.03$, $p = .044$, $\beta_{Area} = -.03$, $t(129) = -0.3$, $p = .770$, consistent with what we found in Experiment 3 and in contrast to the SDT account.

Correlations of Individual Differences. As in Experiments 1 and 3, there was no correlation between children's confidence choices and their Winner, $r(129) = -.01$, $p = .908$, Ask, $r(129) = -.01$, $p = .943$, or Endorse choices $r(129) = .03$, $p = .775$, when controlling for age. We once again conducted a hierarchical regression with age and area accuracy in step 1 and added confidence performance in step 2, again finding no meaningful contribution of either area or confidence performance on the three social judgments.

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 size from Experiment 2, we did not replicate the correlation we previously found between children's sensitivity to confidence and their selective social learning choices.

Mega-Analysis

Our failure to find a consistent correlation between self and other reliability judgements in three out of four experiments we ran suggests that – contrary to our initial predictions – the mechanisms supporting self vs. other judgements are distinct, or at least not following the theorized mechanistic link between SDT and Simulation Theory. At the same time, however, this conclusion relies on a null result. To quantify the strength of this null finding, we perform a mega-analysis, combining the results from the four experiments for maximal power.

To help clarify our findings, and particularly to determine whether we should interpret non-significance as evidence in favor of *no* effect, we computed Bayes Factors (BF). Bayes Factors provide the relative weight of the evidence for the null vs. the alternative hypotheses, and can therefore provide a measure of graded strength for the null hypothesis (Wagenmakers et al., 2018). If neither the null nor alternative hypotheses are supported, the model will output a BF_{10} of 1. If there is support for the alternative hypothesis, values will increase towards positive infinity, and if there is support for the null hypothesis, values will decrease towards 0. All Bayesian analyses were conducted in JASP with default priors.

Selective Social Learning Task

Combining the data from all four studies (N = 374), we find that children aged 5 and up gave Winner and Ask judgments consistent with our predictions and previous work, and children aged 6 and 7 gave consistent Endorse judgments (see Table 5). Thus, largely replicating the

Table 5

Means and Tests Against Chance Across All Experiments.

Age	Mean (%)	SD	*t*	*p*	*d*	BF_{10}
Winner						
4	49.18	24.14	-0.27	.792	0.03	0.15
5	62.38	62.38	4.80	< .001	0.46	2841.68
6	77.43	28.12	9.90	< .001	0.98	$4.46 * 10^{13}$
7	87.86	20.07	19.15	< .001	1.89	$1.77 * 10^{32}$
Ask						
4	51.23	23.46	0.41	.684	0.05	0.15
5	58.88	27.53	3.34	.001	0.32	18.75
6	68.20	31.15	5.93	< .001	0.58	280741.29
7	80.34	24.91	12.36	< .001	1.22	$8.42 * 10^{18}$
Endorse						
4	46.31	24.93	-1.16	.252	0.15	0.26
5	50.23	27.36	0.09	.930	0.01	0.11
6	58.25	29.58	2.83	.006	0.28	4.70
7	58.73	30.60	2.60	.011	0.29	8.74
Confidence						
4	62.65	11.93	8.28	< .001	1.06	$5.61 * 10^{8}$
5	67.63	13.36	13.66	< .001	1.32	$9.48 * 10^{21}$
6	71.68	14.17	15.53	< .001	1.53	$2.66 * 10^{25}$
7	74.45	13.24	18.74	< .001	1.85	$3.25 * 10^{31}$

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, with evidence moderately in favor of the null hypothesis.

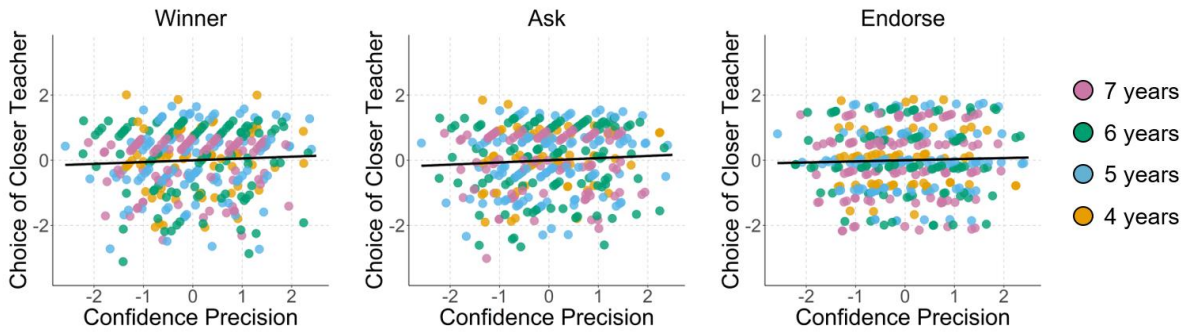
Correlations of Individual Differences

Combining the results from the four studies, there were no correlations between confidence sensitivity and Winner, $r(371) = .06$, $p = .289$, $BF_{10} = .18$, Ask, $r(371) = .07$, $p = .202$, $BF_{10} = .28$, or Endorse choices, $r(371) = .03$, $p = .517$, $BF_{10} = .20$, when controlling for age

(see Figure 2). Importantly, the Bayes Factors were less than 1/3 for including confidence performance as a predictor of the three Selective Social Learning variables, which can be interpreted as moderate evidence for the null hypothesis (i.e., that confidence sensitivity does *not* predict social judgments). By inverting the Bayes Factors, this null hypothesis is 5.71 times more likely than the hypothesis that confidence is a meaningful predictor for Winner judgments, 3.61 times more likely for Ask judgments, and 5.09 times more likely for Endorse judgments.

Figure 2

Correlations Between Confidence and Social Judgments Controlling for Age.



Note: Data plotted are standardized residuals.

General Discussion

We set out to test an account linking reasoning about one's own accuracy and others' accuracy through SDT and Simulation Theory. This account hypothesized that confidence judgments are entirely calculated from decision imprecision, and that social reasoning is accomplished by simulating the confidence judgments of another. In four studies, we presented children with a social learning task and a confidence reasoning task that reduced shared mental state language and response biases, investigating whether imprecision in shared area

representations alone led to correlations between the tasks. However, in three of the four studies, we found no evidence of a correlation between the self and other tasks, and a mega-analysis of the four studies revealed support in favor of *no* correlation. At the same time, both the Selective Social Learning and Confidence Tasks demonstrated sensible patterns replicating past work, indicating that the measures were tapping into the target constructs. Therefore, even under the most generous interpretation of our findings, in which the significant correlation in Experiment 2 represents the true correlation, there is a much weaker correlation than would be expected given the SDT/Simulation account hypotheses. Our interpretation, then, is that reasoning about the likelihood of one's own accuracy and the likelihood of another's accuracy are *not* computed by a single process rooted in representational variability.

Although our findings are not what we expected given previously reported correlations between self and other tasks, they are consistent with other recent evidence against an SDT account of confidence judgments. For example, using the same relative confidence task in 6-9-year-olds, Baer et al. (2018) reported correlations between confidence tasks that use independent representations, which should be uncorrelated according to pure SDT principles (and see Baer & Odic, 2020a). In the Supplemental Material, we partially replicate this finding using data from Experiment 1: area and number confidence judgments were correlated, but neither correlated with emotion confidence. Similarly, Maniscalco and Lau (2012) found that decision imprecision in adults did not perfectly predict confidence ability, again counter to the SDT prediction (see Baer & Odic, 2019 for similar developmental evidence). We replicate this finding in the Supplemental Material using the Confidence Tasks in Experiments 3 and 4, demonstrating that there is age-linked improvement in confidence reasoning beyond what can be explained by change in representational imprecision. Taken together, it seems increasingly unlikely that

representational imprecision alone accounts for metacognitive confidence judgments, against the core SDT predictions.

We will note that the SDT account is still being tested with modifications that we do not account for in the current study. As one example, some SDT theories accommodate a mismatch between confidence and accuracy by inferring that there is additional variability in how well a participant can interpret the imprecision of their decision (e.g., Mamassian, 2020; Maniscalco & Lau, 2012; Rahnev et al., 2011). That is, under this view we should never expect representational imprecision to perfectly explain confidence judgments because there is a secondary process involved (e.g., akin to a statistician who calculates the standard error, but who sometimes presses the wrong number on the calculator or rounds off extra digits). As another example, some SDT theories posit a 'winner-takes-all' computation of confidence, whereby a participant computes confidence not as the difference or ratio between the decision variability, but rather focuses only on the strength of evidence for the chosen answer (e.g., Miyoshi & Lau, 2020; Zawadzka et al., 2017; Zylberberg et al., 2012). In our study, this could involve children tracking the overall statistics about 'average' shape size and inferring the most confidence for the largest shapes, regardless of their size relative to the other shape in the pair. As our study was not designed to test these versions of SDT theories, we cannot say for certain that SDT accounts in general cannot explain confidence judgments or judgments of other's accuracy, only that the particular SDT account tested here does not.[4]

[4] Because we ensured that high confidence trials did not always feature the largest shape, we had those trials in which size was inconsistent with expected confidence (e.g., the largest shape was part of a difficult trial, about 40% of trials). Inconsistent with the winner-takes-all account, children still chose the trial with a higher ratio rather than the one with the largest shape more than chance (p's $< .003$) which may mean that the relative confidence format discourages this heuristic.

However, the lack of correlation alone does not conclusively rule out Simulation Theory as an explanation for reasoning about social judgments. For instance, while we designed our tasks to avoid certain response biases like overconfidence, there are still many potential biases that children could carry that operate in one task but not the other and could therefore serve to mask any underlying correlation. Children could have maximized their success in the Confidence Task as we intended, but instead maximized fairness over success in the Selective Social Learning Task by alternating between informants rather than consistently choosing the more accurate one (e.g., Shaw & Olson, 2012). This pattern would result in accurate measurement of the Confidence Task, but near chance-like performance on the Selective Social Learning Task. In other words, children could still have *understood* which informant was more accurate, but their performance might not have reflected this reasoning. We similarly know that there was some variability in children's motivation in the Confidence Task from the presence of children who consistently chose the harder option, though they may have responded strategically in the Selective Social Learning Task as we anticipated. Though these patterns were likely uncommon given the above-chance performance on both tasks, these potential inconsistent motivations could have weakened any existing correlations. It is thus possible that children simulated their own reasoning when making judgments of others, but the resulting correlation was masked by conflicting motivational goals. Such response strategies would necessarily have to be much stronger than any individual differences in theorized shared mechanisms but are nonetheless plausible.

Alternative Theoretical Accounts

One account that could accommodate our findings through a single common process comes from Bayesian accounts of cognition (Meyniel et al., 2015; Pouget et al., 2016). In most

Bayesian theories, children's decisions are based on weighted evidence, rationally combining prior expectations and available evidence from multiple sources (see Gopnik & Bonawitz, 2015). For this to work, all information must be available in a common unit: in this case, the probability of accuracy. Information like decision imprecision easily complies – more precision tends to indicate a higher probability of accuracy – but so could many other cues, such as one's prior history of success or failure on the task, momentary distraction, mind wandering, and so on. For example, if by chance a teacher always smiles when telling the truth, then smiling could be construed as signalling accuracy. But notice that these same cues in a different context may sometimes signal errors: a teacher smiling during the *child's* statement could reflect an attempt to encourage the child to work through an incorrect answer, or an answer with very high precision (that feels *too* easy) could hint that the question was misinterpreted. Cues are therefore not inherently diagnostic but must be interpreted by the child as meaningful for their decision.

Critically, a Bayesian account allows the child to determine how heavily each cue should be weighed for a given decision. If a child is familiar with 'trick questions' that feel easy but are actually difficult, then that child might not treat precision and the accompanying feelings of ease as a valid cue that they should be confident. In the same way, the cues that children deem relevant for their own decisions and for others' decisions need not be the same. For instance, bodily cues like posture, facial expression, and reaction time may more heavily inform judgments of another's accuracy than one's own, while representational imprecision and feelings of ease may factor more heavily into judgments of one's own accuracy (see Vuillaume et al., 2020). We might therefore not expect a correlation between self and other judgments if children weighed cues differently for the two judgment types, even with a single common Bayesian mechanism. Put differently: our results are consistent with self and other judgements relying on

the same general process, but on distinct *information* (see also Thomas & Jacoby, 2013; Tullis, 2018).

A different account that advocates for a single underlying process of self and other reasoning is the *mindreading-first* account. Essentially the inverse of Simulation Theory, the mindreading-first account argues that all metacognitive reasoning is accomplished by turning one's mindreading abilities inward (e.g., attending to one's own behavioral cues as though observing the self; Carruthers, 2009; Gopnik, 1993). As with Simulation Theory, the lack of correlation between self and other tasks in the current work is inconsistent with this account. However, due to its focus on behavioral cues, the mindreading-first account presents one additional explanation for why we did not detect a correlation. In the Confidence Task, children could have attended to many cues including reaction time or states of anxiety (Carruthers, 2009; Koriat & Ackerman, 2010a; Paulus et al., 2014). However, in the Selective Social Learning Task, children did not see videos or live performances, and so did not have these or many other common behavioral cues that are known to impact children's assessment of accuracy (Birch et al., 2010; Paulus et al., 2014). This imbalance in cues could in part explain why only children aged 5 and older performed above chance on the Selective Social Learning task, while even 4-year-olds succeeded at the Confidence Task where more cues were available. Future work would need to find ways of equating the kinds of cues that children attend to for themselves vs. for others in order to establish such a correlation.

As a final theoretical explanation, these findings are also consistent with theories suggesting full independence of self and other reasoning (e.g., Nichols & Stich, 2003). Though several studies have documented correlations (e.g., Gopnik & Astington, 1988; Kuzyk et al., 2020; Lockl & Schneider, 2007), they did not control for common features that may have

induced correlations. Our experimental design eliminated two such features: common response biases (e.g., to seek as much information as possible) and common language (e.g., using mental state terms like "know"). Consistently, there was no correlation between self and other reasoning in young children in one study where these features were unaligned (Bernard et al., 2015). There, children were asked to opt out of an answer when uncertain (a metacognitive measure without the use of mental state terms and with reward-maximizing) and completed several classic theory of mind tasks (mindreading tasks with mental state terms but no personal investment). Together with the current work, these findings suggest that previously reported correlations could reflect other common features such as response biases or similar demands on mentalistic language, masking the potential independence of self and other reasoning.

Briefly, we will note that correlations driven by response biases or common language may still provide meaningful information about how children *use* self and other judgments. For example, one recent proposal is that response biases are a critical part of how learning occurs, potentially even more than the ability to distinguish between close estimates of accuracy that we isolated here (Baer & Odic, 2020b). That is, learning-relevant behaviors like choosing to seek help or discounting misinformation from others may depend much more heavily on how a child *interprets* their state of confidence (or evaluates whether another person is accurate *enough* to trust). Though our findings could suggest independence in how those signals of self and other accuracy are generated, common response biases gesture towards a single process responsible for *interpreting* those signals. The presence of common linguistic markers ("know," "sure") further supports this idea.

Our findings are also relevant to a recent proposal about the emergence of metacognitive ability during childhood which specifically relies on social learning (Heyes et al., 2020). The

Cultural Origins hypothesis argues that children learn to reason metacognitively by watching others model good metacognition or by having a teacher guide them, rather than metacognition emerging either innately through genetic programming or through non-social experience. A critical prediction of this account is that children with good social skills (and in particular, good social learning skills that help them select the best teachers to learn from) should then show the best metacognitive skills (see Heyes et al., 2020, p. 358). Our findings are inconsistent with this account: children with good selective social learning performance (i.e., those children who selected the best teachers) did not have the best metacognitive skills.

Replication of Sensitivity to Another's Error Magnitude

These results also provide an important replication of the Selective Social Learning 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). Five-year-olds not only noticed the difference in sizes and attributed this to superior ability (by choosing the closer informant as the 'Winner'), 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 Einav and Robinson's (2010) 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, 4-year-olds did not attribute any differences in the magnitude of error to the informants' ability, nor did they selectively Ask or Endorse the closer informant. This is not likely due to difficulty 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 (and not mutually-exclusive) explanations. 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 their effect.

We also found that children in Experiments 1-3 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 Experiment 4 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 overriding a desire to be fair to both informants due to plausible deniability (e.g., Shaw et al., 2014). Future work testing each of these possibilities could be useful not only in understanding children's behavior on selective learning tasks, but more generally in

understanding how children balance epistemic and social goals (Jaswal & Kondrad, 2016; Landrum et al., 2015).

Replication of Relative Confidence Reasoning

We also replicated and extended the findings of Baer and Odic (2019; Baer et al., 2018) in the relative confidence task. As reported in their studies, children responded to their confidence in area discriminations by selecting the easier question (Experiments 1, 2, and 4) or the more accurate answer (Experiment 3). We further replicated their findings of age-related improvement, and in the Supplemental Material show that this is not due to developing area reasoning. Extending this work, we saw that even 4-year-olds were significantly above chance in their confidence reasoning in all four studies. This is currently the youngest age group reported to compare their confidence between two questions, supporting the prediction made by Baer and Odic (2019) that children under age 5 may show sensitivity to confidence if given large enough contrasts in difficulty or an easier task (as the area task is relative to the number task, see Odic, 2018).

Conclusion

Overall, we did not find evidence to support the SDT/Simulation account, and in fact found ample evidence *against* the SDT predictions. Our primary conclusion is therefore that children do not reason about both self and others through a single, representational-imprecision-driven process, as predicted by SDT. These results also point towards no correlation between self and other reasoning, though as outlined, alternative non-SDT or non-simulation-based theoretical accounts could still accommodate this evidence with a single process, which warrants further investigation. However, these experiments *do* support claims that children can reason

strategically about both their own and another's reliability, demonstrating flexibility in their learning mechanisms through tools that help navigate truth from fiction.

References

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

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

Baer, C., & Odic, D. (2019). Certainty in numerical judgments develops independently of the Approximate Number System. *Cognitive Development*, *52*, 100817. https://doi.org/10.1016/j.cogdev.2019.100817

Baer, C., & Odic, D. (2020a). Children flexibly compare their confidence within and across perceptual domains. *Developmental Psychology*, *56*, 2095–2101. https://doi.org/10.1037/dev0001100

Baer, C., & Odic, D. (2020b). The relationship between children's approximate number certainty and symbolic mathematics. *Journal of Numerical Cognition*, *6*(1), 50–65. https://doi.org/10.5964/jnc.v6i1.220

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

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., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science*, *13*(2), 363–369.

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

Brannon, E. M., Lutz, D., & Cordes, S. (2006). The development of area discrimination and its implications for number representation in infancy. *Developmental Science, 9*(6), F59–F64.

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

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

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

Crivello, C., Phillips, S., & Poulin-Dubois, D. (2018). Selective social learning in infancy: Looking for mechanisms. *Developmental Science, 21*(3), e12592. https://doi.org/10.1111/desc.12592

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

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

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning, 10*, 347–374. https://doi.org/10.1007/s11409-014-9133-z

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

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

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

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

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

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.

Hagá, S., & Olson, K. R. (2017). Knowing-it-all but still learning: Perceptions of one's own knowledge and belief revision. *Developmental Psychology, 53,* 2319–2332. https://doi.org/10.1037/dev0000433

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

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

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

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences*, *24*, 349–362. https://doi.org/10.1016/j.tics.2020.02.007

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

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social Metacognition: An Expansionist Review. *Personality and Social Psychology Review*, *2*(2), 137–154. https://doi.org/10.1207/s15327957pspr0202_6

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

Kim, S., Sodian, B., Paulus, M., Senju, A., Okuno, A., Ueno, M., Itakura, S., & Proust, J. (2020). Metacognition and mindreading in young children: A cross-cultural study. *Consciousness and Cognition*, *85*, 103017. https://doi.org/10.1016/j.concog.2020.103017

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science, 15*, 694–698.

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

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., & 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

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

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

Lecce, S., Demicheli, P., Zocchi, S., & Palladino, P. (2015). The origins of children's metamemory: The role of theory of mind. *Journal of Experimental Child Psychology, 131,* 56–72. https://doi.org/10.1016/j.jecp.2014.11.005

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

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, *108*, 732–739. https://doi.org/10.1016/j.cognition.2008.06.013

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. (2007). Knowledge about the mind: Links between theory of mind and later metamemory. *Child Development*, *78*(1), 148–167. https://doi.org/10.1111/j.1467-8624.2007.00990.x

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*, 19751–19756. https://doi.org/10.1073/pnas.0704452104

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development*, *82*, 1778–1787. https://doi.org/10.1111/j.1467-8624.2011.01649.x

Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science*, *2*, 459–481. https://doi.org/10.1146/annurev-vision-111815-114630

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21,* 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology, 24,* 470–476. https://doi.org/10.1037/0012-1649.24.4.470

Meltzoff, A. N. (2007). 'Like me': A foundation for social cognition. *Developmental Science, 10*(1), 126–134. https://doi.org/10.1111/j.1467-7687.2007.00574.x

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

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review, 127*(5), 655–671. https://doi.org/10.1037/rev0000184

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds* (p. 237). Clarendon Press/Oxford University Press. https://doi.org/10.1093/0198236107.001.0001

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., Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Developmental change in the acuity of approximate number and area representations. *Developmental Psychology*, *49*(6), 1103.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258. https://doi.org/10.1126/science.1107621

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

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

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.

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

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

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

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

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

Thomas, R. C., & Jacoby, L. L. (2013). Diminishing adult egocentrism when estimating what others know. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 473–486. https://doi.org/10.1037/a0028883

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

Tullis, J. G. (2018). Predicting others' knowledge: Knowledge estimation as cue utilization. *Memory & Cognition*, *46*(8), 1360–1375. https://doi.org/10.3758/s13421-018-0842-4

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., & van de Pol, J. (2019). Judging own and peer performance when using feedback in elementary school. *Learning and Individual Differences*, *74*, 101754. https://doi.org/10.1016/j.lindif.2019.101754

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

Vuillaume, L., Martin, J.-R., Sackur, J., & Cleeremans, A. (2020). Comparing self- and hetero-metacognition in the absence of verbal communication. *PLoS One, 15*, e0231530. https://doi.org/10.1371/journal.pone.0231530

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

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

Zawadzka, K., Higham, P. A., & Hanczakowski, M. (2017). Confidence in forced-choice recognition: What underlies the ratings? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 552–564. https://doi.org/10.1037/xlm0000321

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience, 6.* https://doi.org/10.3389/fnint.2012.00079