Spring 2023

Children's Intuitive Theories of Group Collaboration

Emory Richardson
Yale University Graduate School of Arts and Sciences, erichardson989@gmail.com

Abstract
Children's Intuitive Theories of Group Collaboration
Emory Richardson
2023

Cumulative culture has amplified human knowledge far beyond what any individual learner could teach themselves or even be taught in a single lifetime: a doctor can learn to drive a car without having to literally reinvent the wheel. In some cases, complex artifacts can be produced through the accumulation of incremental improvements over time; but frequently, they're a product of more or less direct collaboration. But collaboration comes with costs of its own. This thesis is about the cognitive capacities individuals need for collaborative learning to be worth the trouble. I first describe a set of interrelated obstacles to the accumulation of technical knowledge by individual learners capable of learning from each other, and outline some of the tradeoffs of relying on collaborative learning to overcome these obstacles. I then compare these tradeoffs with children's and adults' preferences for more or less direct forms of collaboration. I then focus on how reasoning about speed-accuracy tradeoffs in collaborative and individual learning might constrain those preferences. Finally, I discuss what these studies could tell us about the development of collaborative learning and our understanding of distributed cognitive systems.

Children's Intuitive Theories of Group Collaboration

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

by
Emory Richardson

Dissertation Director: Frank C. Keil

May 2023

Acknowledgements

There are always more people to thank than words to thank them with. To Frank, for his curiosity, intellectual excitement, and his willingness to take a chance on mine. To Mark, for teaching me how to ask kids better questions, interest in discussing the big questions, and his dedication to building collective resources for online research. To Mandy, Flora, Nicole, Aaron, and Sami, for making the lab a fun place. To my dev area cohort, for being there — and to Emily especially, for having her ducks in a row for all of us. To Rosie, for cheerful criticism, methodological insight, and genuine kindness. To Dave Rand, whose understated advice that every incoming grad student should have a Twitter account to follow scientists was worth considerably more than its weight in gold. To Boaz, Sayuri, and Samantha, for helping me find my feet in a new field. To Krishnan, whose advice has never lost its power: doing things that sound interesting will always lead you to interesting people who can introduce you to more interesting things. To Anya and Taya, for support, love, time, being, and everything that words can't express. To my parents, for everything.

Contents

Chapter 1

Introduction

This chapter contains text from the following manuscripts:

Richardson, E., Hok, H., Shaw, A., & Keil, F. C. (*in prep*). Herding cats: Children's intuitive theories of persuasion predict slower collective decisions in larger and more diverse groups, but disregard factional power.

Richardson, E., Davis, I., & Keil, F. C. (*in prep*). Agenda setting and The Emperor's New Clothes: People infer that letting powerful agents make their opinion known early can trigger information cascades and pluralistic ignorance.

Richardson, E., & Keil, F. C. (2022). Anger, evidence, & trending opinions: We trust consensus when we believe it reflects genuine persuasion. *PsyArXiv*.

Richardson, E., Miro-Rivera, D., & Keil, F. C. (2022). Know your network: People infer cultural drift from network structure, and expect collaborating with more distant experts to improve innovation, but collaborating with network-neighbors to improve memory. *Proceedings of the Cognitive Science Society, 44*.

Cumulative culture has made us into obligatory social learners. In extreme cases, this is self-evident: a single human life is simply not long enough for an isolated autodidact to teach themselves even a fraction of what they need to know to send a rocket to the moon — good luck inventing calculus from scratch, much less microchips, metalworking, mining, fuel production, and so on. Fortunately, firsthand understanding usually isn't needed even when it's possible; it can be outsourced to the community. Cognitive labor, like physical labor, is divided; and the division of labor allows communities to develop and maintain specialized knowledge that individuals simply wouldn't have the time to acquire even if they had the capacity. But as specialization increases, collaborations have to scale up as well in order to keep pace. We not only need enough collaborators with the right kinds of knowledge; we need collaborators who are capable of *using* each other's knowledge in a timely manner. And while the *necessity* of collaboration becomes increasingly apparent as cultural knowledge accumulates, the cognitive capacities that *enable* us to collaborate are less obvious.

1.1 What's missing in the study of the cognitive capacities that enable collaborative learning?

The lack of clarity is not for lack of study. In many respects, collaborative learning is just a species of social learning with some cooperation thrown in — and cognitive scientists have studied both for decades (Laland, 2004; Harris, Koenig, Corriveau, Jaswal, 2018; Rand, 2016; Whiten, 2017; Moll & Tomasello, 2007; Köymen & Tomasello, 2020). Even as children, we prefer to cooperate with people who have demonstrated prosocial behavior, adherence to in-group norms, and good (or at least, rational) judgment. We trust informants who are kind, competent, and have relevant expertise. And we use a variety of heuristics to infer whether our partners have these traits, as well as which traits matter most in a given setting. But, collaboration can't be easily reduced to a kind of cooperative social learning (Tomasello, Melis, Tennie, Wyman, & Hermann, 2012). The

lack of clarity about what enables us to collaborate so successfully is due to the problem being genuinely tricky. Having some examples on the table will help illustrate.

- The Inherited Errors Problem. Learning from Bob lets Alice avoid the costs of learning alone, and lets her learn things she wouldn't have been able to learn through firsthand trial-and-error; but she also risks inheriting Bob's errors. And the errors may not even be his. Bob may have learned from Carol, who learned from David, and so on. In other words, social learning is a double-edged sword: it allows *errors* to cascade through long transmission chains as well as knowledge (Boyd & Richerson, 1995). Social learners have to be vigilant to avoid inheriting each other's errors. And even as children, we are. But no single person has the time or ability to evaluate everything on the merits, particularly as cultural knowledge accumulates and specialization increases.

- The Heuristics Problem. Heuristics about individuals' competence, cooperativeness, and access to information can serve as evidence-by-proxy of their reliability as informants. But heuristics are a gambit: experts make mistakes, friends lack expertise, and eyewitnesses blink and lie — the humdrum contingencies of life can distort the testimony of a hundred informants in just as many ways, and the most reliable informants won't necessarily be willing or able to cooperate. And no single heuristic is a panacea anyway; they need to be applied flexibly to work at all. For instance, if Bob is kind but incompetent and Carol is competent but unkind, Alice needs to decide which trait matters more (Danovitch & Keil, 2007). The Inherited Errors Problem adds to the challenge. If Alice hears from Bob after Bob hears from Carol, heuristics about their individual traits won't help Alice at all unless she can use them to evaluate Carol's influence on Bob's reliability. The division of cognitive labor forces us to rely on others, but often leaves us with collaborators whose knowledge is second- or third-hand at best.

- The Mutual Influence Problem. If transmission errors make social learning a double-edged sword, there's a sense in which collaborating directly simply sharpens the blade. Like any kind of social learning, collaboration can improve judgment or harm it. For instance, Bob may point out that David has made a mistake; neither Alice nor David can see why, but Carol realizes Bob is right, and Alice and David defer to her expert explanation. But suppose David was right after all, and Carol was simply hoodwinked by Bob's confidence; Alice and David may still defer to Carol's expertise-infused explanation without recognizing her lapse in judgment. For that matter, even if they all *know* Bob is wrong, they might just not think he's wrong in a way that's worth fighting over — but they may or may not be right about how much it matters. These kinds of social dynamics can exacerbate both the Inherited Errors Problem and the Heuristics Problem. Collective errors can be inherited just like individual errors, but (1) collective errors are endorsed by multiple informants instead of just one, and (2) since allowing individuals to influence each other means that the reliability of their collective judgment

> no longer supervenes on their reliabilities *as individuals*, heuristics about individual reliabilities will be misleading unless you know how the individuals have influenced each other (List & Pettit, 2006; Goldman, 2014; Dunn, 2019). Intuitively: if the group decision was driven by Bob the Blowhard, Carol's expertise doesn't matter — and vice versa. As a downstream learner, you can save time and resources by relying on collective judgments; but if you don't know enough to evaluate a judgment on the merits, you ignore mutual influences between your informants at your peril. What cumulative culture highlights is that there are almost always mutual influences upstream of all us, whether we're standing on the cliched shoulders of scientific giants, doomscrolling through viral promotions from consolidated media companies and troll factories, or listening to one of our collaborators convincing another to change her mind about a project decision.

Here's the gist. Since no one can learn everything on their own, our learning capacity depends on how much we can rely on what we learn from others. But *their* learning capacities do too. And since we're all downstream of *some* mutual influences between our informants, we not only need to consider those influences when evaluating each others' beliefs — we need to be able to count on our informants to do the same. This dissertation is about some of the commonsense intuitions that could allow us to do that even in childhood, particularly in collaborative contexts. I focus on collaborative contexts because (for reasons I hope will become clear as we go) it seemed like the most familiar setting for the kinds of commonsense intuitions I have mind. These intuitions are about (1) whether mutual influence is more likely to make consensus judgments more reliable or less, and (2) how speed-accuracy tradeoffs affect individual and collective judgment. Equipped with roughly accurate intuitions about how these factors constrain the judgment of individuals and the communities of learners they collaborate with, learners may be able to make better use of that judgment without being misled by it. They may also be better equipped to collaborate in ways that make the benefits of collaboration worth the time and effort.

The plan for what follows goes like this. In the remainder of Chapter 1, I give an overview of the theoretical background to the work as a whole. I'll first synthesize several areas of literature concerning the strengths and weaknesses of collaborative learning in groups. Each section in Chapter 1 reviews research with both adults and children; the penultimate section summarizes the potential causes of developmental change and how studies of children's metacognitive development can shed light on adult capacities. In Chapter 2, I contrast two kinds of social learning strategies (small group discussion versus crowdsourcing) that vary in the opportunity informants have to influence each other, for two kinds of questions (demonstrative reasoning versus non-reasoning) that vary in the risks and benefits of enabling that influence. In Chapters 3 and 4, I examine commonsense reasoning about the speed-accuracy tradeoffs of individual (Chapter 3) and collective (Chapter 4) judgment. Chapter 3 is motivated by the observation that while *thinking takes time*, perception and memory can be accurate on much faster timescales. Chapter 4 is motivated by the observation that while collaborators have to cede unilateral control over speed-accuracy tradeoffs, they may still be able to influence how the group manages those

tradeoffs by reasoning about three endogenous constraints on group decision speed — the number of people, the number of factions, and the (im)balance of factional power. All experiments in these chapters are pre-registered. In Chapter 5, I begin by synthesizing the results of the preceding chapters and discussing potential limitations and extensions, with a particular focus on implications for theories of conceptual development. I then discuss how intuitive theories of group collaboration could shed light on our capacity for cumulative culture and the implications for recent debates about cognitive systems at the group level.

1.1.2 (Reasoning about) consensus: a double-edged sword

Consensus often has a strong influence on individual and group judgments — not only in humans (across ages, cultures, and a variety of contexts), but in other species as well (Morgan & Laland, 2012; Morgan, Rendell, Ehn, Hoppitt, & Laland, 2012; Haun, van Leeuwen, & Edelson, 2013; van Leeuwen et al., 2018; Boehm, 1996).

In some respects, this is unsurprising: formal models originating with Condorcet and Galton demonstrate that as long as informants' judgments are statistically independent from each other, increasing the number of informants makes the strength of consensus increasingly likely to reflect the average competence of the crowd (Condorcet, 1994/1785; Dietrich & Spiekermann, 2013; List & Goodin, 2001). Roughly speaking: consensus in a competent crowd amplifies accurate judgments, but consensus in an incompetent crowd amplifies inaccurate judgments, simply because uncorrelated errors cancel each other out. For instance, even if the average informant only answers a yes-no question accurately 60% of the time, a large enough crowd all but guarantees that the judgment of the *majority* will be accurate. Consensus' reliability and the relatively minimal cognitive capacities required to compute it would seem to make it a good learning strategy. And it's made consensus a widely studied topic across the social and evolutionary sciences (Hastie & Kameda, 2005; Conradt & List, 2009; Centola, 2022; Kameda, Toyokawa, & Tindale, 2022).

But in other respects, reliance on consensus is very surprising: Condorcet's logic assumes individuals' judgments are statistically independent. And since social learning is ubiquitous in the biological world, that's a notoriously implausible assumption to make (Laan, Madirolas, de Polavieja, 2017; Kao & Couzin, 2014). Even shared perceptual and cognitive biases are sufficient to compromise statistical independence, to say nothing of shared culture, environment, or motivated reasoning. There's no guarantee that these biases are truth-conducive, and widespread reliance on consensus-based learning strategies can create feedback loops that amplify them even further (Hahn, von Sydow, Merdes, 2020; Becker, Almaatouq, & Horvat, 2020; Almaatouq, Rahimian, Burton, & Alhajri, 2020).

Perhaps learners (and humans in particular) are just particularly careful about avoiding dependencies to the extent possible? Evidence is mixed at best. Even when presented with an explicit choice between a consensus with a high potential for dependencies and one with low potential, many people explicitly defend their decision to trust the former more than the latter (Xie & Hayes, 2022; Yousif, Aboody, & Keil, 2019, Exp. 4). But sensitivity to "false" consensus also seems to differ by context. For instance, in an eyewitness memory context, adults and children as

young as six trust one person reporting the testimony of four eyewitnesses more than four people reporting the (opposite) testimony of one eyewitness (Aboody, Yousif, Sheskin, & Keil, 2022; Yousif, Aboody, & Keil, 2019, Exp. 5). And even preschoolers reject a "false consensus" when experimenters make clear that some informants' judgments are unjustified, such as by presenting a consensus that repeats the testimony of an informant who was "pretending" to have seen the contents of a box, or a consensus that disagrees with an artist about whether she drew a basketball or an orange (Kim & Spelke, 2020; Einav, 2014). In contrast, adults give equal weight to an economic forecast that appears in five news articles citing a single primary source and one that appears in a single article citing five primary sources (Yousif, Aboody, & Keil, Exp. 1-3). And when three informants all endorse the same answer to a trivia question, five year olds trust the consensus *more* when two informants ostentatiously copy the first's testimony than when each informant answers independently — only by age 8-9 do children show the adult pattern (Einav, 2018). Why? One reason might be that the benefits of being able to learn from others even when their judgments aren't independent are worth the risks. For instance, people might expect each others' judgment to be helped by direct communication more than it's hurt, at least in some contexts. But is it?

1.1.3 (Reasoning about) the scope of potential benefits in group collaboration

The intuition is literally proverbial: *two heads are better than one*. The literature on problem-solving in groups has frequently supported folk wisdom on this point. For example, individual accuracy on the original Wason card selection task is typically around 10%; however, when undergrads are prompted to discuss the problem as a group and decide on a consensus answer, the accuracy rate rises to 75%, *even when no group member's initial solution is correct* (Moshman & Geil, 1998). Small discussion groups also outperform even the best individuals on a variety of inductive reasoning tasks (Laughlin et al., 2006; Laughlin, 2011; Trouche, Sander, & Mercier, 2014), concept learning in undergraduate genetics courses (Smith et al., 2009), lie detection (Klein & Epley, 2015), jury decisions (Guarnaschelli et al., 2000), and numerical estimation (Navajas et al., 2018). Groups even punch above their weight relative to large crowds: in one study, averaging the numerical estimates of four discussion groups produced a more accurate answer than even 1400 individuals (Navajas et al., 2018). What allows groups to outperform crowds, and even their best members?

1.1.3a The benefits

A now-common view is that groups are better at processing information than individuals (Hinsz, Tindale, Vollrath, 1997). There are several possible reasons for this. For instance, even though managing a discussion is cognitively demanding, group discussion may, on the whole, reduce cognitive load on individual members for complex problems (Laughlin, 2011). If each member focuses on partially overlapping areas of same complex problem, the group as a whole can increase its "collective working memory" without requiring any individual member to consider every aspect of the problem (Kirschner, Paas, & Kirschner, 2009; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Similarly, complex problems frequently require expertise

from diverse domains. To the extent that a society divides cognitive and physical labor, individuals are unlikely to all have expertise in every domain they need in order to solve a problem, but groups have a better chance. Another potential advantage is that group discussion often takes the form of collective explanation. Even explaining to *oneself* is an effective learning tool; compared to participants asked to merely describe or think aloud, children and adults asked to explain causal phenomena to themselves are more likely to notice inconsistent evidence, outright mistakes, and gaps in their own understanding (Chi, De Leeuw, Chiu, & Lavancher, 1994; Teasley, 1995; Williams & Lombrozo, 2010). As a result, they achieve deeper levels of understanding and better integrate new information with prior beliefs than non-explainers (Walker, Lombrozo, Legare, Gopnik, 2014; Walker, Lombrozo, Williams, Rafferty, Gopnik, 2017; Legare & Lombrozo, 2014). Group discussion may amplify the benefits of self-explanation by allowing collaborators to notice *each others*' inconsistencies, mistakes, and gaps as well (Chi, Roy, & Hausmann, 2008). Moreover, needing to resolve differing perspectives into a single solution may motivate collaborators to find those gaps in others' explanations as well as clarify their own (Mercier & Sperber, 2011).

For instance, when Pine & Messer (1998) classified 5-7-year-old children according to their pre-tested intuitive theories of balance, placed them in "same-theory" or "different-theories" 4-person groups, and gave each child beams to balance on fulcrums, children who had been asked to discuss with their groups improved more from pre-test to post-test than those who had not. Moreover, the improvement was greater for different-theories groups than same-theory groups, and the different-theory groups produced more than three times as many "challenges" to each other's explanations. In short, even in early childhood, groups may be more reliable and efficient than individuals because they improve our ability to pool evidence, generate solutions, and catch mistakes. By adulthood, these advantages allow groups to solve a variety of complex constraint satisfaction problems more quickly and accurately than any individual (Laughlin, 2011; Almaatouq, Alsobay, Yin, & Watts, 2021).

1.1.3b The scope

But the advantages of group discussion may not apply equally to every problem learners encounter, or every group they form — for children or adults (Sears & Reagin, 2013; Nokes-Malach, Richey, & Gadgil, 2015; Nokes-Malach, Meade, & Morrow, 2012; Williams, Lombrozo, & Rehder, 2013). For instance, Mercier & Claidière (2022) asked fair-goers to each think silently for 5 minutes about one of three kinds of problems (ethical, trivia, and "demonstrable" deductions) and then discuss it in small groups for 10 minutes; they then elicited minute-by-minute changes in each participant's beliefs. Responses to ethical questions (e.g., dollar value of compensation for losing a finger in a workplace accident or finding worms in your soup) didn't change one way or the other over time. Responses to trivia questions (e.g., the total goals scored in the 2010 World Cup or the total number of elevators in the Empire State Building) improvements in accuracy only reached significance after discussion and the patterns of improvement varied between groups and questions. However, questions with *demonstrably* correct answers (e.g., the bat-and-ball or Paul & Linda problem) made for a fairly dramatic contrast with those patterns. After

discussion, responses were almost unanimously correct in all groups even though fewer than 20% of responses were initially accurate, and thinking silently only led to much smaller improvements among a small subset of people.

What makes the outcomes for "demonstrable" questions so different from ethical and trivia questions? Demonstrability is not an intrinsic feature of tasks themselves; it's a matter of the resources that group members have available for *accurately adjudicating* each other's judgments — in time, motivation, task-relevant information, and competence in shared conceptual systems. But some kinds of tasks offer more resources for adjudication than others. And that degree of "demonstrability" seems predict the relationship between a group's final decision and the number of votes that decision initially received (Laughlin & Ellis, 1986; for review, see Bonner, Shannahan, Bain, Coll, & Meikle, 2021; Kerr & Tindale, 2004).

For instance, the shared conceptual systems people use to evaluate ethical judgments and memory for trivia simply don't afford "proofs" of accuracy; and since groups default to majority rule (Boehm, 1996; Laughlin, 2011), the answer most people initially believe is the answer the group will typically endorse, unless people change their minds for other reasons. But the system we use to adjudicate logical and physical reasoning does afford more conclusive proof — for groups that are sufficiently competent to evaluate each other's use of it. Even if most people can't prove that squares are half the area of the square on their diagonal, small discussion groups will adopt the right answer as long as a single member finds the solution at some point during the discussion — in short, "truth wins", even the if a majority initially endorsed the wrong answer (Laughlin & Ellis, 1986; Mercier & Claidière, 2022).

Don't mistake the point here. These shared conceptual systems don't have to make individuals better at solving problems *for themselves (*though they might, especially if they motivate people to spend more time and effort). The point is that as long as they make people better at evaluating *someone else's* judgment, they can make group judgments more accurate. But in many domains, the conceptual competencies that enable us to accurately adjudicate others' judgments seem to include both learned skills and innate capacities. Mathematics may be "the preeminent domain of demonstrability" (Laughlin & Ellis, 1986), but for young children, the demonstrability of doubling a square may be no higher than deducing the speed of light would be for the average adult (Socrates' demonstration to Meno's slave boy notwithstanding). And since adjudicating disagreement through demonstrative reasoning takes more time and effort than simpler strategies (e.g., deferring to confidence, prestige, or consensus), collaborators may be less motivated to spend time and effort deliberating until they understand what can be gained from it. Developmental research on demonstrability may shed light on how we acquire an understanding of the costs and benefits of deliberation. As we'll see in Chapter 2, these tradeoffs suggest that the more value people see in demonstrative reasoning, the more they'll see the value of collaboration as outweighing the risks.

Of course, there may be other reasons to prefer collaboration; for instance, discussion may facilitate coordination on various kinds of arbitrary norms. Though this dissertation focuses on high-demonstrability contexts, I'll briefly discuss the potential nuances of extending these ideas to other contexts in Chapter 5. For instance, people may be less concerned that discussion will

"bias" decisions about arbitrary norms than whether the consequences of allowing conventions to emerge ad hoc would be costly enough to justify spending time to coordinate in advance (e.g., driving vs. walking on the same side of the road).

1.1.4 (Reasoning about) costs, risks, and challenges in group collaboration — and how to minimize them

Groups obviously aren't infallible. Groupthink and conformity à la Asch and Milgram are familiar risks to psychologists — though often overstated: sensationalism in textbooks and undergraduate lectures notwithstanding, most people in Asch's experiments did not conform most of the time. Across trials, 25% of participants never conformed at all, and with even a single "ally", conformity fell from a high of 35% per trial to only 5%, (Asch, 1956; Griggs, 2015). Even among children, conformist tendencies are larger for matters of convention than moral or perceptual judgments, and children privately reject incorrect majorities even when they publicly conform (Haun & Tomasello, 2011; Pham & Buchsbaum, 2020). And strong egocentric biases emerge in a variety of contexts: we often give less weight to others' judgment than we should (Mannes, 2009; Morgan et al., 2015), are skeptical of both expert consensus and people we suspect are blindly conformist, and offer more- and less-nuanced reasons for that skepticism (Light et al., 2022; Oktar & Lombrozo, 2022).

But even when collaborators are *able* to avoid the pitfalls of conformist tendencies, it may not always be worth the costs in time, effort, and social cohesion. Collectively, the members of a team may *have* all the skills and information they need to improve their solution to a problem through intensive collaboration (e.g., discussion). But they can't *use* each others' knowledge unless they're aware of it. And learners need time, effort, and fairly sophisticated cognitive capacities to discover what their collaborators know and coordinate joint attention on the relevant aspects of a problem (Laughlin, 2011). The higher these (expected) costs go, the the greater the margin by which the (expected) payoffs of collaboration will have to exceed those of strategies with lower upfront costs (Almaatouq, Alsobay, Yin, & Watts, 2021), such as estimating an initial consensus and using it to make a decision without further ado. Consider some examples of challenges intrinsic to collaboration, whose costs groups have to pay upfront in order to make use of their collective capacities.

1.1.4a Hidden Profiles

Laboratory studies of problem-solving groups suggest that by age 9, we're much more likely to discuss mutually-shared information than information known only to one member — their "hidden profiles" (Stasser & Titus, 2003; Gummerum, Leman, & Hollins, 2014; Ker & Tindale, 2004). As a result, groups often underperform the sum of their members' knowledge. In some respects, finding that ad hoc groups in psychology labs overlook hidden profiles seems analogous to demonstrating that small-talk at cocktail parties is shallow and general; obviously, (WEIRD) strangers will quickly seek common ground, and are unlikely to ask each other probing questions without special motivation. If the task doesn't *necessitate* (or at least *motivate*) contributions from every teammate, individuals are more likely to free-ride, and their team may underperform the

sum of its members' knowledge simply because that much of that knowledge never came to light. This is not to say that hidden profiles aren't a real problem; it's not always clear how much effort any given task is worth or what you need to know to solve it (Stasser & Abele, 2020). But some tasks make that easier to see than others. Give the cocktail party a problem with a demonstrably correct answer, and they'll do better: when led to believe that a murder mystery included sufficient evidence to deduce the culprit, groups were twice as likely to identify the culprit, and brought up the "hidden profile" clues for discussion 20% more often than groups led to believe that the evidence was inconclusive (Stasser & Stewart, 1992). The point is that even when uncovering hidden profiles *can* improve collective judgment, collaborators need to have the skills, motivation, and opportunity to do so.

1.1.4b Joint attention

Collaborators can't simply attend to a problem themselves; they have to ensure they're all attending to the same problem. Joint attention can be as simple as the mutual awareness of looking at the same physical object, which emerges in pre-verbal infants around 9 months (Abney, Suanda, Smith, & Yu, 2020). But we can also attend to the same *mental* objects. The problem is that doing so is more cognitively demanding and often considerably more time-consuming (Wohltjen & Wheatley, 2021). Consider the example of Sally, who has read that the moon is made of cheese:

Sally: "It must be cheese! It's yellow, and cheese is yellow!"

Mother: "*That* doesn't mean it's made of cheese — your socks are yellow, and they're not cheese!"

O'Madagain & Tomasello (2019) point out that in Mother's utterance, the discourse demonstrative *that* refers "not just a belief, but a *reason* for holding a belief". In other words, Mother is able to convince Sally that the moon is not cheese because they share conceptual systems (e.g., verbal reasoning & assumptions about substance-color-form relationships) that allow them to jointly attend to *mental* objects like reasons as well as physical objects like socks.

But reasons are complex relationships between conceptual theories, observable evidence, and the access reasoners have to them. Without a shared conceptual system that enables collaborators to keep track of each others' beliefs and the reasons for holding (or changing) them, collaboration may not be much use even for experts. For instance, if one expert is reporting a nuclear emergency to another, these systems for sharing reasons allow him to condense the evidence, conclusion, and implications of a reactor core explosion into a single phrase: "*there's graphite on the ground*". But in explaining the situation to a bureaucrat, the expert would need to clarify that in a nuclear plant, graphite is only found in the core of reactor; thus, if there are chunks of graphite on the ground outside it, the core must have exploded with enough force to eject the graphite; thus, such-and-such consequences are already inevitable, and such-and-such actions are urgently necessary. In order to benefit from attending to each others' reasoning, collaborators not only need to be competent enough in the relevant conceptual systems that they're able to help each

other understand the evidence being presented — they need to have the motivation and skill to do so efficiently (Bonner, Shannahan, Bain, Coll, & Meikle, 2021).

The ability to reason from common ground appears quite early; like a nuclear technician explaining or merely implying the significance of graphite depending on the expertise of their interlocutor, even three-year-olds are more likely to justify claims to partners who don't share common ground, and omit those justifications when they're part of common ground (Köymen, Mammen, & Tomasello, 2016). But more broadly, sharing reasons with collaborators also means both *providing* and *evaluating* approximate explanations of matters we only partially understand to begin with (Keil, 2006). Children are much less skilled than adults in adjudicating conflicting explanations, and many of the skills we use to more efficiently evaluate each others' reasoning — such as engaging in meta-talk comparing higher-order evidence such as our relative confidence or their informants' reliability — only begin to emerge between the ages of five and seven (Köymen & Tomasello, 2018). The point is that our ability to co-construct explanations and adjudicate between them is constrained by our shared conceptual systems and the capacity for joint attention we use to navigate them. But these activities are cognitively and socially demanding even when done well; doing them poorly makes them even more costly. And doing them well gets harder as collaborations (and the problems they address) grow larger and more complex (Cooney, Mastroianni, Abi-Esber, Brooks, 2020). If Alice zones out while working alone, she can simply resume her task whenever she likes; and in a dyad with Bob, her lapse may be salient enough that he calls her out right away; but larger groups can make lapses of attention more likely to occur (since keeping track of each others' reasoning is more difficult), less likely to stand out in the crowd, and harder to compensate for once conversation has moved on.

1.1.4c Time management

Conversation moves quickly even when it seems to take forever to get where it's going (Templeton, Chang, Reynolds, LeBeaumont, & Wheatley, 2022; Mastroianni, Gilbert, Cooney, & Wilson, 2021). But the decisions it leads to and the time it takes to arrive are constrained by factors that are only partially and imperfectly under our control. Consider some examples. Bob may have a chance to shape the conversation by jumping in with a quick answer before others have had the time to think it through; his collaborators will then have to decide whether to accept his proposal and move the discussion forward from there, spend more time deliberating, or redirect entirely. If the group realizes Alice has zoned out or misunderstood, they can decide whether her informed input is critical enough to repeat what's been said. If David disagrees with an emerging consensus, he'll have to decide whether it's worth fighting over; and if he seems intent on filibustering, his collaborators will have to decide whether his willingness to spend the time and effort is more indicative of bullheadedness or an insight they've overlooked. Deciding to collaborate means that even though no one has unilateral control over either the group's decision or decision speed, both are constrained by each person's individual decisions about how to interact with each other. The more judiciously we pick our battles and maintain collective focus, the more the benefits of collaboration will outweigh the costs.

In Chapters 2 and 4, I review evidence suggesting that the influence individuals' "patience" has on speed-accuracy tradeoffs in individual and collective judgment isn't specific to discussion, or even species (Albrecht, Anderson, & Vroman, 2010; Chan, Lizzeri, Suen, & Yariv, 2018; Sumpter & Pratt, 2009; Sasaki, Stott, & Pratt, 2019). But to my knowledge, there's relatively little work on how children reason about time management. Preschoolers do expect more difficult physical tasks to take longer for individual agents (Leonard, Bennett-Pierre, & Gweon, 2019); and by middle childhood, individual students' performance on Raven's Progressive Matrices is strongly correlated with how much they modulate the time they spend on easy versus difficult items (Perret & Dauvier, 2018). Moreover, reasoning about physical costs constrains our interpretations of each other's actions even in early infancy (Liu, Ullman, Tenenbaum & Spelke, 2017; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). But time costs are intrinsic to every cognitive process as well as physical action, making them more ubiquitous than physical costs. Research on the developmental origins of how we reason about them could shed light on a fundamental constraint on human cognition.

1.2 Developmental origins and mechanisms

The experiments in this dissertation won't be able to speak to the ontogenetic *origins* of our collaborative learning skills at either the biological or cognitive level. But the developmental trajectory of children's reasoning about the tradeoffs of individual and collective deliberation may be able to shed light on the *constraints* that shape our capacity for collaborative learning even as adults. Aside from maturational changes in attention and memory, what might those constraints be?

One constraint may be theory of mind development. The kind of first-order theory of mind reasoning that allows us to evaluate Bob's beliefs about the world has traditionally been said to emerge between the ages of 3 and 5 (Wellman, 2001), although many of the components likely emerge earlier (Scott & Baillergeon, 2017). But reasoning about how Bob's beliefs about the world influence Alice's beliefs may require higher-order theory of mind, which has been said to emerge between the ages of 5 and 7 (Miller, 2009). The challenge of higher-order theory of mind reasoning help explain why Mills, Al-Jabari, & Archacki (2012) found that only 21% of 6-year-olds spontaneously mention bias in explaining why a judge's decision to award first prize in a contest to their friend might elicit protest, or why Einav (2018) found that before age 8-9, children trusted consensus *more* when two informants ostentatiously copied a third's testimony than when each informant answered independently. Detecting a mismatch between the world and Bob's beliefs about it only requires a representation of one mind; estimating the influence of inherited errors may require representing many more minds, making phenomena like bias, groupthink, and hearsay more difficult for young children to evaluate — particularly for judgments that aren't inherently implausible or intended to deceive (Cottrell, Torres, Harris, & Ronfard, 2023; Mills, 2013). The assumption that our informants are exercising some degree of epistemic vigilance and a capacity for rational belief updating can make us particularly vulnerable to inherited errors; after all, the more people endorsing a claim, the harder it is to explain away the consensus as a failure of vigilance or rationality. Revealing informants' evidence and their reasons and for

believing it can make evaluating their claims easier for young children; but real-world bias, error, and deception usually aren't announced in advance. I'll return to these points in Chapter 5.

A second constraint may be our familiarity with culturally inherited conceptual systems for evaluating claims. While the elements of logic may be untaught knowledge — even infants engage in process-of-elimination reasoning (e.g., a cup contains either A or B; but since it's not B, it must be A; Cesana-Arlotti, Martín, Téglás, Vorobyova, Cetnarski, & Bonatti, 2018) — one of the larger themes in this dissertation is our extreme and ubiquitous dependence on specialized expertise. Even in pure mathematics, the elements of logic don't buy us logical omniscience (Hardwig, 1991). Understanding phenomena like thermal expansion and gene expression requires expert assistance, and without them it may be difficult to understand why people would be convinced by Richard Feynman's explanation of the *Challenger* explosion or biologists' confidence in evolutionary theory. However, abstract intuitions about which kinds of arguments are most compelling — to ourselves and others — may help children learn to evaluate each others' beliefs as well as their own as they incorporate culturally acquired conceptual systems into their beliefs (Lombrozo, 2016). This would also be congruent with recent accounts emphasizing that reasoning itself is often a social process, and best thought of as a set of constraints on our intuitive evaluations of what constitutes a "good reason". Accounts differ on whether these intuitions originate in genetically endowed modules (Mercier & Sperber, 2011; 2019) or culturally inherited cognitive gadgets (Heyes, 2019; Novaes, 2020). Evidence of early emerging intuitions about which kinds of problems enable more reliable demonstrations of accuracy may further our understanding of reasoning's social functions. I return to these points in Chapter 5.

1.3 An overview of the upcoming chapters

In Chapter 2, I focus on how reasoning about the risks and benefits of mutual influence may prompt different social learning strategies for different kinds of problems. For instance, small group discussions allow individuals to pool their knowledge, process evidence collectively, and point out mistakes; but discussion can also exacerbate social biases and groupthink, and conversation itself can become unmanageable as the group grows larger. Many of these problems can be avoided by simply polling a crowd for individual answers; but while crowds can be surprisingly wise, they may also amplify the kinds of biases and mistakes that a discussion could have easily corrected. So when might we expect groups outperform crowds? Across three studies, I find that for "reasoning" questions, adults and children as young as six prefer to expect small group discussion to be more helpful than either asking each to answer alone or polling a crowd of ten times as many individuals. However, they show the opposite pattern for advice on challenging perceptual discrimination tasks and inferences about population preferences.

In Chapter 3, I suggest that reasoning about the speed-accuracy tradeoffs of an agent's judgment and decision-making processes can help us evaluate the reliability of their beliefs and make sense of their behavior. The core insight is that while *thinking takes time*, perception and memory can be accurate on much faster timescales. This means that responses that are too-fast or too-slow give us a lot of leverage for explaining people's beliefs and behaviors. A moment's

hesitation when asked whether the ball you're holding is pink or red might suggest the color is ambiguous; the same hesitation might raise suspicion if you're asked whether you were at the bank during the robbery; but it might imply that you're only offering a rough estimate instead of a precise calculation when asked how much gas money an electric car saved you this year. Across three studies, I find precisely these kinds of inferences in adults and children as young as six. Moreover, children expect the amount of time agents spend thinking about a task to depend on the complexity of the task.

Chapter 4 explores a conceptual link between Chapters 2 and 3. Collaboration allows us to solve problems we might not be able to solve alone, but it also comes with costs — we may need to recruit collaborators, agree on a joint goal and how to accomplish it, and maintain enough social cohesion in the group to do so. Since all of these things takes time and effort, the benefits of collaboration may not always be worth the cost. For instance, groups commonly make decisions by consensus. You may be sure of the best way for the group to accomplish a goal, but find that a strong majority favors an alternative approach. You could try to convince them otherwise, but would it be worth the time and effort? It would help to know how much time that would take — but you would need to have some way to estimate those time costs. I suggest that people may be able to estimate the time needed to reach consensus by reasoning about three endogenous constraints on group decisions — the number of people, the number of factions, and the (im)balance of factional power. In two experiments, I find that adults and children as young as six expect slower decisions from teams with more people or more factions; but only adults expect faster decisions from teams with stronger initial consensus. Reasoning about how factional power constrains group decision speed may help make us more efficient collaborators.

Chapter 2

Groups and crowds

This chapter is based on materials published in Richardson, E., & Keil, F. C. (2022). The potential for effective reasoning guides children's preference for small group discussion over crowdsourcing. *Scientific Reports*, *12*(1), 1193. https://doi.org/10.1038/s41598-021-04680-z

Abstract

Communication between social learners can make a group collectively "wiser" than any individual, but conformist tendencies can also distort collective judgment. We asked whether intuitions about when communication is likely to improve or distort collective judgment could allow social learners take advantage of the benefits of communication while minimizing the risks. In three experiments (n=360), 7- to 10-year old children and adults decided whether to refer a question to a small group for discussion or "crowdsource" independent judgments from individual advisors. For problems affording the kind of 'demonstrative' reasoning that allows a group member to reliably correct even errors made by a majority, all ages preferred to consult the discussion group, even compared to a crowd ten times as large — consistent with past research suggesting that discussion groups regularly outperform even their best members for reasoning problems. In contrast, we observed a consistent developmental shift towards crowdsourcing independent judgments when reasoning by itself was insufficient to conclusively answer a question. Results suggest sophisticated intuitions about the nature of social influence and collective intelligence may guide our social learning strategies from early in development.

2.1 Introduction

When is advice from multiple people more likely to clarify than confound a learner's understanding? Consider two ways one could learn from multiple people at once: by eliciting a consensus judgment from a small group discussion, or by "crowdsourcing" many independent answers. Discussion may enable groups to correct mistakes and combine insights, producing an accurate consensus answer that no individual could have found alone. However, without an objective method of evaluating solutions, discussion may drag on endlessly, be misled by charismatic leaders or groupthink, and ultimately only create an illusion of consensus around a wrong answer. In contrast, crowdsourcing may still include mistakes that discussion could have corrected, but particularly in *large* crowds of *independent* responders, the majority, plurality, and average response can all be surprisingly accurate (Hastie & Kameda, 2005, Laan, Madirolas, & de Polavieja, 2017). Nevertheless, shared culture and cognitive biases can create illusions of consensus even without direct communication between individuals [Yousif, Aboody, & Keil, 2019; Mercier & Miton, 2019). The empirical advantages of discussion and various crowdsourcing strategies are well-documented. Less attention has been given to laypeople's own intuitions about the tradeoffs between them (note that our use of "intuition" does not refer to the intuitive-deliberative distinction in dual systems theory; rather, it follows the frequent use of "intuitive theories" in developmental psychology to describe the untaught assumptions about the world that help learners structure their experience; for a recent review, see Gerstenberg & Tenenbaum, 2017). Here, we investigate early developing intuitions about when group discussion or crowdsourcing is a more effective use of collective intelligence.

While debate over whether crowds can be trusted is at least as old as philosophy itself, mathematical models suggest that under certain conditions, crowds can be "wise". Given a set of options to "vote" for, majority and plurality accuracy increase to near certainty as crowd size increases (List & Goodin, 2001; Dietrich & Spiekermann, 2013), and evolutionary simulations suggest that conformist learning strategies are often more adaptive than alternatives [Boyd & Richerson, 1988; Hastie & Kameda, 2005). Similarly, averaging crowd members' individual judgments can produce a collective estimate that is more accurate than the crowd's most accurate member (Galton, 1907; Hong & Page, 2004; Steyvers, Miller, Hemmer, & Lee, 2009; de Oliveira & Nisbett, 2018; Laan et al., 2017). Interestingly, many species faced with the problem of learning from multiple sources at once rely on similar heuristics to evaluate collective opinion. Even in early childhood, people trust majority over minority judgment, and give more weight to stronger majorities (Morgan, Laland, & Harris, 2015). By adulthood, people also trust pluralities, and give more weight to the judgments of larger crowds [Muthukrishna, Morgan, Henrich, 2016; Mannes, 2009; Morgan, Rendell, Ehn, Hoppitt, & Laland, 2012).

However, crowdsourcing heuristics share a common weakness: for crowds to be "wise", individual judgments must be independent. Social influence can compound individual error, particularly when a large proportion of the population are conformist learners [Raafat, Chater, & Frith, 2009; Lorenz, Rauhut, Schweitzer, & Helbing, 2011). Yet, while popular concerns about echo chambers and media bias suggest that laypeople intuitively recognize some of the risks of social influences, it remains unclear how well people compensate for them in practice. For example,

while adults and even children as young as six prefer firsthand knowledge over hearsay in an eyewitness memory context [Yousif et al., 2019; Aboody, Yousif, Sheskin, & Keil, 2019), adults are just as trusting of an economic forecast repeated by five news articles citing a single primary source as they are of the same forecast citing five different primary sources (Yousif et al., 2019; Sulik, Bahrami, & DeRoy, 2020). Similarly, while children as young as four expect randomly sampled evidence to cause others to revise their beliefs more than evidence from a biased sampling process (Magid, Yan, Siegel, Tenenbaum, & Schulz, 2018), when the source of the sampling bias is *the selection of informants itself*, children are sometimes insensitive to bias even late in development— particularly when the degree of consensus is high (Whalen, Griffiths, & Buchsbaum, 2018; Anderson & Holt, 1997; Einav, 2018; Hu, Whalen, Buchsbaum & Griffiths, 2015; Mills & Keil, 2005; Mills & Grant, 2009). Indeed, even as adults, people frequently mistake the frequency of a belief in their local networks for its frequency in the population as a whole [Marks & Miller, 1987; Lerman, Yan, & Wu, 2016; Stewart eta l., 2019). In short, people's trust and distrust of consensus seems to selectively disregard one of the proposed preconditions of consensus' accuracy — independent sources.

One reason for people's occasional indifference to their sources' independence may be that social influence frequently makes their judgments *more* accurate (Becker, Brackbill, & Centola, 2017; Becker, Porter, & Centola, 2019; Abel & Bauml, 2020; Mason & Watts, 2012; Derex & Boyd, 2015; Barkoczi & Galesic, 2016). For instance, while open discussion may risk groupthink by sacrificing individuals' independence, it also allows individuals to pool their knowledge and generate new insights; discussion can also ease the cognitive load on individuals, increase a groups' capacity for processing information, and allow the group to correct individual mistakes (Kirschner, Paas, & Kirschner, 2009a; Kirschner, Paas, & Kirschner, 2009b; Laughlin, 2011). This division of cognitive labor means that discussion groups may be able to quickly generate solutions that most individuals would never produce alone, and may make discussion an attractive learning strategy for a wide variety of problems, particularly as the evidence load increases (Smith et al., 2009; Laughlin, Bonner, & Altermatt, 1998; Laughlin, Bonner, & Miner, 2002; Almaatouq, Alsobay, Yin, & Watts, 2021). Most notably, to the extent that discussion enables even a single group member to correct a *majority* that has made a mistake, discussion may also have an advantage over crowdsourcing heuristics like majority rule (Moshman & Geil, 1998). Studies of group problem-solving have suggested that this 'truth wins' effect occurs when a shared conceptual system enables individuals to conclusively *demonstrate* that a given answer is correct or incorrect — and it is the strength of their argument, rather than the individual's confidence or simply the presentation of the correct answer, which predicts whether the majority will be persuaded. Importantly, these studies suggest that "demonstrability" is a matter of degree, ranging from mathematics as the "preeminent domain of demonstrability" to purely judgmental tasks such as attitudes and preferences, with a variety of evidence-based reasoning and insight problems also being high in "demonstrability" (Trouche, Sander, & Mercier, 2014; Laughlin & Ellis, 1986; Larson, 2010). Note the implication of the "truth wins" effect for social learners: if a minority is able to demonstrate that their judgment is accurate, the majority is not simply *influenced* by the judgment of the minority, they will *defer* to it. Thus, when

demonstrations are possible, discussion groups may offer substantially more accurate collective judgments than a "crowdsourced" majority, with little risk of distorting an accurate majority judgment. Indeed, recent accounts suggest that reasoning itself may be most naturally deployed in service of argumentation and function most effectively in interpersonal contexts [Mercier, 2016; Mercier & Sperber, 2011).

Note that we are not claiming that group discussion is *only* beneficial for questions that afford demonstrative reasoning, or that demonstration and reasoning are synonymous. Rather, we focus on discussion and crowdsourcing as flexible, commonsense approaches to a fundamental problem for any social learner: integrating information from multiple sources without inheriting their errors. Our intent is to examine laypeople's intuitions about their tradeoffs. Though crowdsourcing heuristics like majority rule can be remarkably accurate, they also presuppose independent judges — an unrealistic assumption about human societies. Meanwhile, work on group problem-solving has repeatedly found that discussion not only allows groups to outperform heuristics like majority rule, but that their ability to do so depends on the "demonstrability" of the problem (Bonner, et al., 2021). Of course, demonstration is possible without reasoning (e.g., by physically demonstrating how an artifact works or showing the location of an object), and reasoning cannot always conclusively demonstrate a that a solution is optimal. However, reasoning may be a reliable means of correcting errors even when physical demonstrations are not feasible, and when a correct answer cannot be simply deduced. For example, knowing the distance from New York to Chicago won't allow a group to deduce the distance from New York to Cleveland, but it may enable them correct some over- and underestimates without needing to actually measure the distance. Indeed, in a recent comparison of group discussion with the wisdom of crowds on a numerical estimation task, the average collective estimate of four small-group discussions was more accurate than the average of 1,400 individual estimates, and participants reported arriving at their estimates by "sharing arguments and reasoning together" (Navajas et al., 2018). In short, to the extent that people expect to be able to rely on demonstrative reasoning to minimize the risks groupthink, it may be intuitive to disregard the importance of independent judgment, even if they favor crowdsourcing heuristics in other cases.

2.2 General method

In the present work, we asked whether people would favor different social learning strategies for problems that afford demonstrative reasoning than those that do not. Crowdsourcing independent judgments may be more valuable when the potential for reasoning is less salient, particularly when the crowd is large. Discussion may be more valuable when demonstrative reasoning provides a reliable means of analyzing problems and identifying errors, even if the discussion group is small. Because past work suggests that sophisticated social learning strategies emerge in early childhood but also that children appear to underestimate some risks of social influence even in late childhood (Aboody et al., 2019, Einav, 2018), we focused on adults and children ages 7-10. Understanding how the ability to balance the risks and benefits of social influence develops could shed light on the incongruence of our remarkable capacity for collective

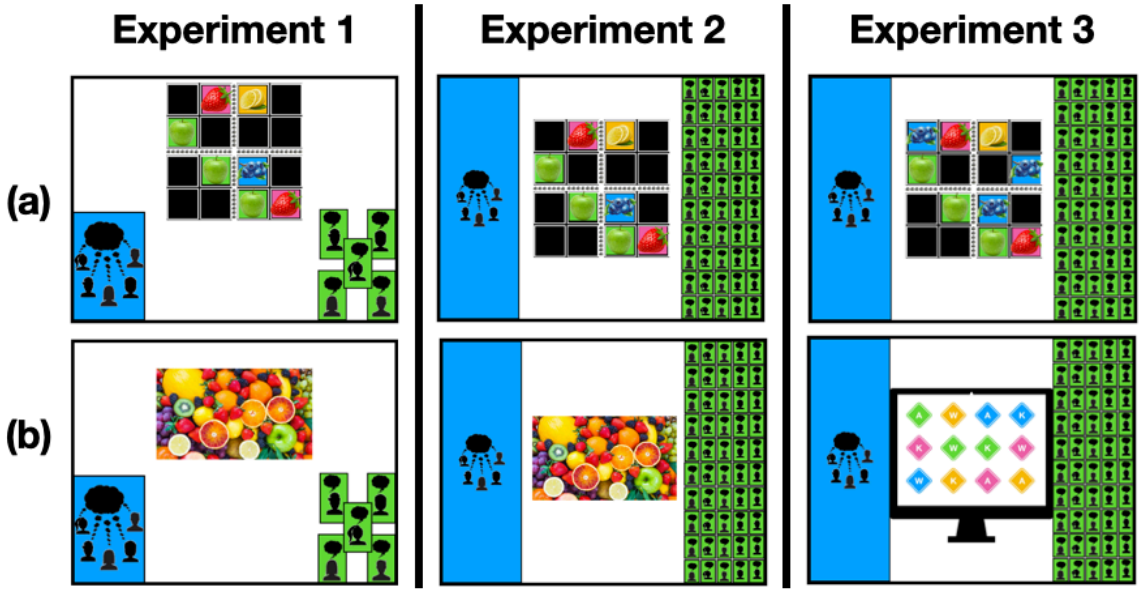


Figure 2.1. Example stimuli: Groups and Crowds**.** Each participant saw 8 questions. For each question, they were asked which way of answering would be more helpful: *"five people talking together"*, or *"[five / fifty] people answering alone"*. Questions were identical in Experiments 1-2 (4 *Reasoning*, 4 *Population Preference*). Experiment 3 contrasted *Easy* versions of the *Reasoning* questions with *Hard Perceptual Discrimination* questions. **(a) Top Row**: Example *Reasoning* question. Two-by-two Fruit Sudoku from Experiments 1 & 2 & partially completed "Easy" version in Experiment 3. **(b) Bottom Row**: Example *Population Preference* question (most popular fruit in the world) and *Perceptual Discrimination* question (which shape is spinning the fastest).

problem-solving and our apparent susceptibility to groupthink. It may also provide clues as to where interventions to thwart misinformation may be most effective.

In each experiment (Figure 2.1), participants were shown eight questions (4 *Reasoning* and 4 *Non-Reasoning*), and for each question, they rated whether crowdsourcing or discussion would be more helpful in answering, on a 4-point scale. In Experiment 1, this meant that participants rated whether it would be more helpful to ask five people to each answer independently, or to ask the same five people to give a single group answer after discussing. In Experiments 2 and 3, we contrasted the five-person group discussion with a crowd of 50 people answering alone. For the *Reasoning* questions, we chose a set of constraint-satisfaction problems that would challenge adults' capacities, but still be understandable to children (e.g., Sudoku). Because the solutions to these questions must satisfy a mutually understood set of explicit constraints, discussion can help groups generate potential solutions and and reduce processing demands on individuals while relying on demonstrative reasoning to correct errors. In Experiments 1 and 2, we contrasted the *Reasoning* questions with *Population Preference* questions (e.g., most popular fruit in the world). Though individuals' intuitions may sway as the discussion generates potential answers, discussion provides no objective means of adjudicating disagreement; thus, it may distort

intuitions rather than sharpen them. In Experiment 3, we contrasted easy versions of the same *Reasoning* questions with a set of challenging *Perceptual Discrimination* questions (e.g., fastest rotating item in an array), which a separate sample had rated as more difficult than the *Reasoning* questions. This allowed us to test the role of perceived difficulty against the potential for effective reasoning. If participants simply favor discussion for questions that feel more difficult — regardless of whether discussion can reliably adjudicate disagreement — then the preference for group discussion will be stronger for the *Perceptual Discrimination* questions than the *Reasoning* questions. Our general prediction in all three experiments was that sensitivity to the contrast between reasoning and intuitive judgment would lead all ages to prefer group discussion for reasoning questions. However, because past work has suggested that children may underestimate the risks of social influence until between the ages of 6 and 9 (Aboody, et al., 2019; Magid et al., 2018; Whalen, Griffiths, & Buchsbaum, 2018; Einav, 2018), we predicted that a robust preference for crowdsourcing non-reasoning questions would emerge only among older children (ages 9-10) and adults, while younger children (ages 7-8) would favor group discussion for both kinds of questions in Experiments 1 and 3. All experiments were preregistered, and the data, materials, and power analyses are available on the OSF repository (https://osf.io/6pw5n/). All experiments were approved by the Yale University Institutional Review Board and conducted according to their guidelines. Written informed consent was obtained from all adult participants. Because children participated online, parents were recorded reading the informed consent form aloud.

2.3 Experiment 1

2.3.1 Method

Participants. We recruited 40 adults through MTurk, as well as 80 children (40 Younger, M=8.01, SD=.56; 40 Older, M=9.92, SD=.56; 39 girls). Children participated through an online platform for developmental research that allows researchers to video chat with families using pictures and videos on slides (Sheskin & Keil, 2018). Sample size was chosen based on the estimated effect size from pilot results.

Materials. We asked eight test questions (Figure 2.1), four from each of two question types: *Reasoning* and *Popularity*. Questions were presented from the perspective of a protagonist (Jack). The *Reasoning* questions were chosen to be simple enough to explain to children, but challenging enough that the answer would not be immediately obvious to adults. (1) A 4x4 Sudoku puzzle adapted for children. (2) A vehicle routing problem that required a MarioKart character to find the shortest road through all the treasures on a map without taking "two in a row that are the same color, or two in a row that are the same shape". (3) A single-heap game of Nim ("Each side takes turns picking up pencils. Each turn, you can pick up either 1, 2, or 3 pencils. The winner is the person who picks up the last pencil. There are 5 pencils left in this game; how many pencils should Jack pick up?"). (4) An "impossible object" puzzle that requires the solver to remove a dowel held in place by a nut and bolt from inside a bottle without breaking the dowel or the bottle. The *Popularity* questions concerned the most common subjective preferences in a population. (1) Whether pizza or hot dogs were more preferred by students in Jack's school. (2)

What most people in the world say their favorite fruit is. (3) What most people in the world say their favorite day is. (4) What most people at Jack's school say their favorite color is. Questions were written to have approximately equal word counts (M_{Reas} = 73.75, M_{Pop}=67.75). Three counterbalances were created to vary the order of the questions — Forward, Reverse, and a Shuffle. Color coding of answer choice and left/right presentation were also counterbalanced between participants.

Procedure. Children were introduced to the protagonist, Jack (a silhouette). They were told that Jack was unsure of the answers to the questions, and could ask five people for help. The five people could either help by *Talking Together* (giving Jack a single answer as a group), or by *Answering Alone* (each giving Jack their own answer after thinking about the question without consulting others). For each item, children and adults rated whether "talking together" or "answering alone" was "probably more helpful, or definitely more helpful", producing a 4-point scale of relative preference, where 1 corresponds to "definitely answering alone", and 4 corresponds to "definitely talking together." Adults used the scale directly; children's responses were staggered: they first chose the more helpful strategy, and then were asked for a "probably/definitely" judgment. After answering the eight test items, participants were asked the two comprehension check questions (these were not counterbalanced: Comp_TT was always presented first). Two features of the procedure are important to keep in mind. First, participants could not evaluate the content of any answer to any question, because none was given: they were asked to choose a *means* of advice, not evaluate the quality of the advice itself. Secondly, they could not make judgments based on *degree* or *quality* of consensus — they only knew that the group would have to give one answer, while the crowd would have to give 5 independent answers which could differ or not.

2.3.2 Results

Results. For the primary test, the four responses within each question domain (Figure 2.2) were averaged to create a single score for each domain. A repeated measures ANOVA revealed a significant effect of question *Type* ($F(1,117)=132.87$, $p<.001$, $\eta_p^2 = .532$) and an *AgeGroup*Type* interaction ($F(2,117)=7.83$, $p<.001$, $\eta_p^2 = .118$), and a marginal but non-significant effect of *AgeGroup* ($F(2,117)=2.82$, $p=.064$, $\eta_p^2 = .046$). Multiple comparisons suggested that intuitions about how to manage collective wisdom appear by at least age 7: consistent with the empirical literature suggesting that group reasoning outperforms individual reasoning, all age groups believed that *Talking Together* would be more helpful than *Answering Alone* for *Reasoning* questions, both as compared to *Popularity* questions (Bonferroni corrected, Younger: $t(117) = 3.66$ $p=0.0057$, Older: $t(117)= 7.105$, $p<.0001$, Adult: $t(117) = 9.201$, $p<.0001$), and compared to chance (Younger: M=3.11, SD=.52, $t(39) = 7.42$, $p<.0001$, Older: M=3.19, SD=.51, $t(39) = 8.53$, $p<.0001$, Adult: M=3.45, SD=.54, $t(39) = 11.237$, $p<.0001$). Moreover, both Older children and Adults favored *Answering Alone* over *Talking Together* for *Popularity* questions, though Younger children's answers for *Popularity* did not differ significantly from chance (Younger: M=2.46, SD=.85 $t(39) = -0.232$, p=n.s., Older: M=1.94, SD=.86, $t(39) = -4.076$, $p<.001$, Adult: M=1.83, SD=.81, $t(39) = -5.24$, $p<.0001$). The preference for group reasoning did not differ by age (all ps >.4), though Older

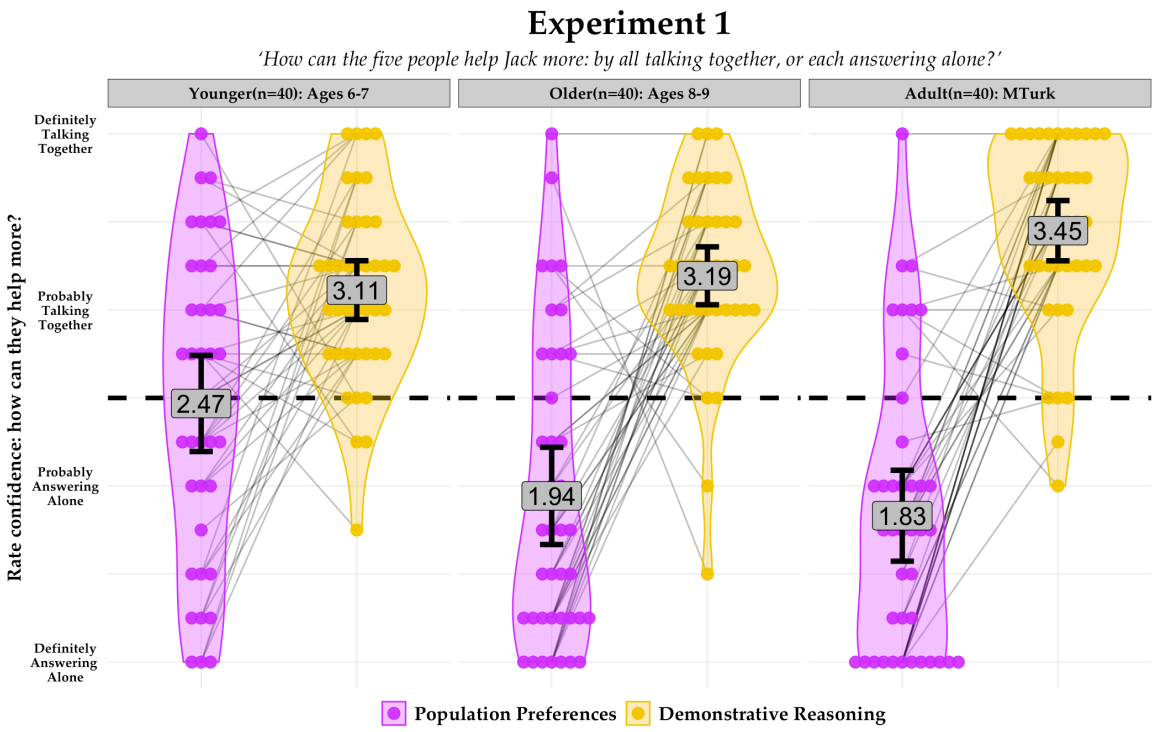


Figure 2.2. Preference for group discussion or crowdsourcing in Experiment 1; each participant's responses were averaged across four *Reasoning* questions (yellow) and four *Population Preference* questions (magenta). Higher ratings indicate stronger preference for group discussion. Grey labels show means; errors bars are 95% CIs; black lines show within-subject differences for the average rating by question *Type*.

children and Adults showed a stronger preference for crowdsourcing *Popularity* questions than Younger children (Bonferroni corrected: Adult vs. Older: $t(78)=0.718$, $p=ns$; Adult vs. Younger: $t(78)=4.067$, $p<0.001$; Older vs. Younger: $t(78)=3.350$, $p<0.0143$). This developmental shift towards *Answering Alone* when discussion provides no objective criteria for evaluating accuracy is slightly earlier than we had predicted, but consistent with past work on children's evaluation of non-independent testimony (Einav, 2018).

Finally, all ages agreed that a teacher who wanted a group of five students to answer test questions accurately should have the students *Talk Together*, while a teacher who wanted to know which students had done their homework should have the students *Answer Alone* (Comp_AA: $M_{Young}= 65\%$, $p=.04$, $M_{Old}= 87.5\%$, $p<.0001$, $M_{Adult}= 92.5\%$, $p<.0001$, Comp_TT: $M_{Young}= 70\%$, $p=.008$, $M_{Old}= 85\%$, $p<.0001$, $M_{Adult}= 87.5\%$, $p<.0001$). This suggests that by age 7, children recognize that discussion could undermine inferences about individuals' "independent" beliefs, but expect group discussion to either generate or disseminate accurate answers.

Taken together, these two tasks suggest that sophisticated intuitions about the risks and benefits of social influence may guide decisions about how to learn from collective judgment. Notably, these intuitions are consistent with empirical findings documenting the a group advantage over individuals for reasoning questions, and the value of independent responding when discussion is likely to bias collective judgment.

2.4 Experiment 2

2.4.1 Method

Could Experiment 1 have underestimated the value of crowdsourcing? Crowdsourcing may be most valuable with large crowds: larger crowds are more likely to include at least one accurate individual, and better represent the relative frequency of beliefs in the population. Moreover, in large enough crowds, even a minimal plurality will easily outnumber the unanimous consensus of a small group. Thus, if a belief's frequency is a cue to its accuracy, a large crowd will always be more informative than a small group. In Experiment 2, we contrasted the 5-person group with a larger 50-person crowd. We predicted that since the *Popularity* questions simply ask the group or crowd to estimate what *most* people in a population prefer, all age groups would find it intuitive to ask *more* people — i.e., the crowd. The benefit of large crowds is less clear for *Reasoning* questions. If few individuals can solve a problem alone, identifying the correct answer in the crowd may be akin to finding a needle in a haystack; indeed, if individual accuracy is known to be rare, the most common answer may be a widely-shared misconception (Prelec, Seung, & McCoy, 2017). Yet, if many individuals can solve the problem alone, large crowds are redundant and a learner can outsource evaluating accuracy to a group discussion. We therefore predicted that adults and older children would continue to favor group deliberation over crowdsourcing for *Reasoning* questions. However, we saw two plausible alternatives for younger children. First, younger children could show the mature pattern. Alternatively, younger children's preference for reasoning in groups could be attenuated by a "more is better" bias. Additionally, since the only difference between Experiments 1 and 2 was the increased crowd size, our design also allows us to explore the effects of crowd size itself by comparing the two experiments directly.

Participants. We recruited 40 adults through mTurk, as well as 80 children (40 Younger, M=8.01, SD=.56; 40 Older, M=9.92, SD=.56; 39 girls). As in Experiment 1, children participated through an online platform for developmental research that allows researchers to video chat with families using pictures and videos on slides (Sheskin & Keil, 2018). One additional child was excluded and replaced because the family lost internet connection partway through the experiment and could not rejoin.

Materials & Procedure. The materials and procedure were identical to Experiment 1, but participants were first shown a large crowd of people, and told that Jack could either ask 5 of them to Talk Together, or 50 of them to answer alone. The answer choices from Experiment 1 were altered to display fifty cartoon icons for *Answering Alone* instead of five.

2.4.2 Results

Results. As before, the four responses for each question *Type* (Figure 2.3) were averaged to create a single score for each *Type*. A repeated measures ANOVA revealed a significant effect of *Type* ($F(1,117)=376.88$, $p<.001$, $\eta_p^2 = .763$) and *AgeGroup* ($F(2,117)=9.63$, $p<.001$ $\eta_p^2 = .141$), and an *AgeGroup*Type* interaction ($F(2,117)=5.39$, $p<.01$, $\eta_p^2 = .084$). Despite the crowd having ten times as many sources as the group, participants were not swayed by a "more is better" bias; all age groups continued to prefer the group discussion for *Reasoning* questions, both as compared to

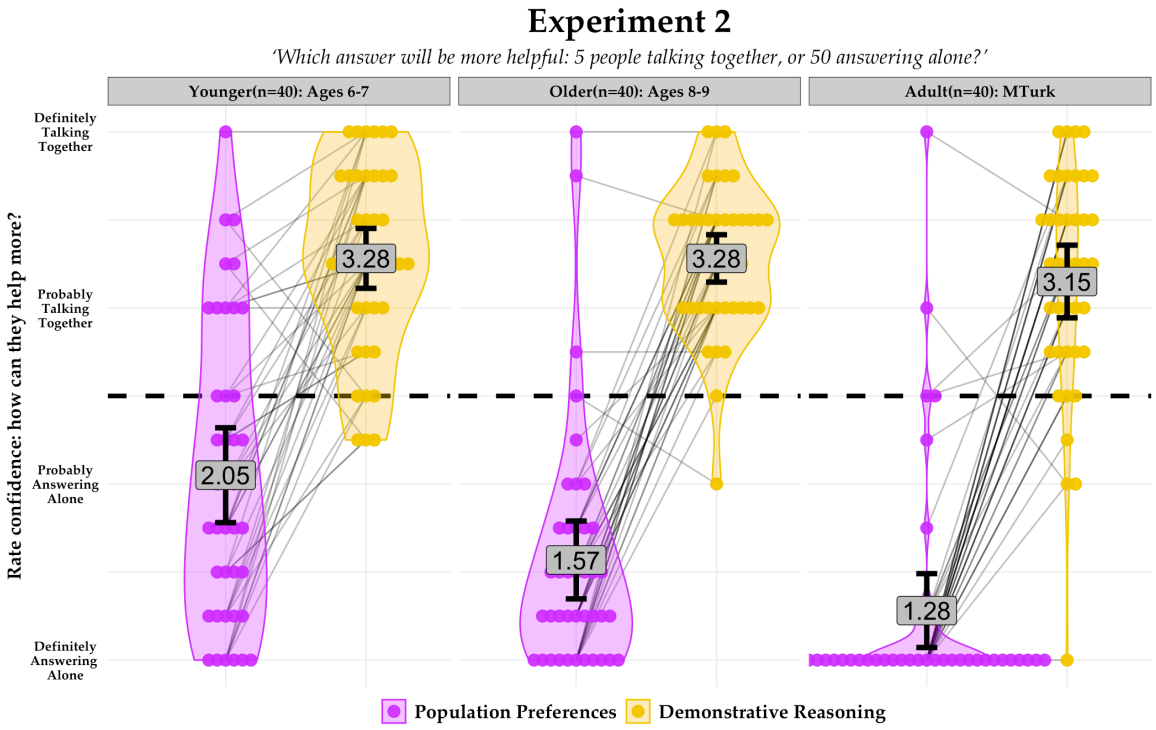


Figure 2.3. Preference for group discussion or crowdsourcing in Experiment 2; each participant's responses were averaged across four *Reasoning* questions (yellow) and four *Population Preference* questions (magenta). Higher ratings indicate stronger preference for group discussion. Grey labels show means; errors bars are 95% CIs; black lines show within-subject differences for the average rating by question *Type*.

Popularity questions (Bonferroni corrected, Younger: $t(117) = 8.60$ $p<.0001$, Older: $t(117)= 11.97$, $p<.0001$, Adult: $t(117) = 13.06$, $p<.0001$), and compared to chance responding (Younger: M=3.28, SD=.53, $t(39) = 9.29$, $p<.0001$, Older: M=3.28, SD=.42, $t(39) = 11.75$, $p<.0001$, Adult: M=3.15, SD=.65, $t(39) = 6.37$, $p<.0001$). Moreover, even younger children in Experiment 2 favored *Answering Alone* for *Popularity* questions, suggesting that they recognized that a large crowd would provide a better estimate of population preferences than a small group (Younger: M=2.05, SD=.84, $t(39) = -3.38$, $p=.0017$, Older: M=1.57, SD=.69, $t(39) = -8.52$, $p<.0001$, Adult: M=1.28, SD=.66, $t(39) = -11.75$, $p<.0001$). As in Experiment 1, the preference for group reasoning did not differ by age (all *p*s > .9), though Older children and Adults again showed a stronger preference for crowdsourcing *Popularity* questions than Younger children (Bonferroni corrected, Adult vs. Older: $t(78)=1.995$, $p=ns$; Adult vs. Younger: $t(78)=5.334$, $p<0.001$; Older vs. Younger: $t(78)= 3.339$, $p< 0.0147$). We also conducted two exploratory analyses of the effect of crowd size. Our preregistered prediction in Experiment 2 was that participants would favor the crowd for population preference questions, but continue to favor the group for reasoning questions. However, because the only difference between Experiments 1 and 2 was the increase in crowd size from 5 to 50 people, our data also enables us to test the crowd-size effect directly. We ran separate ANOVAs for each *QuestionType* using *AgeGroup* & *Experiment* as predictors. The tenfold increase in crowd size had no impact on participants' preference for discussing *Reasoning* questions in small groups ($F(1, 234)=0.045$, $p=.8320$); an *AgeGroup*ExpNum* interaction was

significant ($F(2,234)=4.434$, $p=.0129$), but post-hoc comparisons revealed only a marginal difference between younger children's and adults' preferences for reasoning in groups in Exp 1, with no other differences. However, participants were significantly more likely to crowdsource *Popularity* questions in Experiment 2 than Experiment 1 ($F(1, 234)=19.303$, $p<0.0001$), with no differences between age groups.

As in Experiment 1, responses to the comprehension questions at the end of the task suggested even the youngest children recognized that talking together would make it impossible for the teacher to know which students had done their homework (Comp_AA: $M_{Young}= 67.5\%$, $p=.019$, $M_{Old}= 92.5\%$, $p<.0001$, $M_{Adult}= 90\%$, $p<.0001$). However, while older children and adults agreed that the students would do better on the test if they could discuss their answers, younger children were at chance (Comp_TT: $M_{Young}= 52.5\%$, $p=.4373$, $M_{Old}= 90\%$, $p<.0001$, $M_{Adult}= 90\%$, $p<.0001$). Younger children may be less confident in the value of discussion than their responses to the main task questions in Experiments 1 and 2 would suggest; however, informal questioning of participants after the experiment suggested that younger children in Experiment 2 may have simply rejected talking together on a test as cheating, even though the question specified that the teacher could choose to allow students to talk together.

In short, Experiment 2 suggests that not only are young children's intuitions about the value of group discussion consistent with empirical demonstrations of a group advantage for reasoning questions and the value of large crowds for intuitive estimations. Moreover, directly comparing Experiments 1 and 2 suggests that while children's preference for reasoning in small groups is stable even in the face of a much larger crowd, they also recognize that for some questions, larger crowds are more helpful than smaller crowds.

2.5 Experiment 3

2.5.1 Method

Using *Population Preferences* as the Non-Reasoning questions in Experiments 1 and 2 leaves two points unclear. First, since a culture's preferences are intuitive for most people, the *Popularity* questions may have simply seemed easier to answer than the *Reasoning* questions. Second, because individual preferences are literally constitutive of the population preference, children's responses could reflect an understanding of the nature of preference polling as much as an understanding of the potential for groupthink. To test these two alternatives, we contrasted easy versions of the reasoning questions with challenging perceptual discrimination questions. Disagreement about a challenging perceptual discrimination task would leave a group of laypeople little to discuss beyond confidence, which may be sufficient to filter out obviously wrong answers (Masoni & Roux, 2017; Bahrami et al., 2010), but is generally an unreliable proxy for accuracy. In contrast, polling a large crowd has been shown to increase the accuracy of a collective decision for perceptual tasks (Juni & Eckstein, 2018). The relative difficulty of the *Easy Reasoning* and *Hard Percept* items was confirmed in a pre-test (Supplemental Materials). If participants' preference for group discussion in Experiments 1 and 2 was driven by perceived question difficulty, then they will prefer group discussion for *Hard Percept* questions more than for *Easy Reasoning* questions. If group discussion was preferred because of its perceived benefits

for reasoning, participants will prefer group discussion more for *Easy Reasoning* than *Hard Percept*. If participants recognize the risks of social influence when discussants cannot rely on demonstrative reasoning, they will prefer crowdsourcing for *Hard Percept* questions. We predicted that adults and older children would recognize the tradeoffs, but because children under 8 frequently fail to recognize the potential for motivational biases even in simpler cases (Mills & Keil, 2008), we predicted that younger children would prefer group discussion for both question types. As an exploratory analysis, we also compare the results in Experiment 3 directly to Experiment 2, but given that both the perceived difficulty and the subtype of Non-Reasoning question differ between Experiments, direct comparisons should be interpreted with caution.

Participants. We recruited 40 adults through mTurk, as well as 80 children (40 Younger, M=8.01, SD=.62; 40 Older, M=10.00, SD=.53; 37 girls). As in Experiments 1 & 2, children participated through our online platform (Sheskin & Keil, 2018). Two children were excluded and replaced when database records identified them afterwards as having already participated in Experiment 2. Two adults were excluded and replaced as well; though our preregistered plan was to accept all mTurkers who passed the basic attention screening, two worker identification codes appeared multiple times in the data, passing the attention screen after failing and being screened out two and three times, respectively, in violation of mTurk policies.

Materials & Procedure. Methods were identical to Experiment 2, with the exception of the following changes made to the questions themselves. First, we presented four new Non-Reasoning questions, replacing the four *Popularity* questions with four *Percept* questions: (1) decide which of two pictures of a face "at the tipping point of animacy" is a photo and which is a photorealistic drawing (Looser & Wheatley, 2010), (2) decide whether an opaque box contains 30 or 40 marbles by listening to a recording of it being shaken (Siegel, Magid, Tenenbaum, & Schulz, 2014), (3) identify which of twelve colored squares in a visual array is rotating the fastest, and (4) rank the 25 brightest stars in a photo of the night sky in order of brightness. Second, we simplified the four *Reasoning* questions (see Supplemental Materials) by (1) completing most of the Sudoku, (2) reducing the number of treasures Mario was required to pick up in the vehicle routing problem, and (3) replacing the "impossible object bottle" with an analog of the "floating peanut" task, which requires the learner to extract an object from a jar of water without touching the jar or object (Hanus, Mendes, Tennie, & Call, 2011). The fourth Reasoning question, Nim, remained the same, as adults rated the 5-item Nim heap as easy to solve.

2.5.2 Results

Results. For the primary test, the four responses within each question domain (Figure 2.4) were again averaged to create a single score for each *Type*. A repeated measures ANOVA again revealed a significant effect of question *Type* ($F(1,117)=56.12$, $p<.0001$, $\eta_p^2 = .324$) and *AgeGroup* ($F(1,117)=7.01$, $p=.0012$, $\eta_p^2 = .108$) an *AgeGroup*Type* interaction ($F(2,117)=4.40$, $p=.0143$, $\eta_p^2 = .070$). The perceived difficulty of the questions had no discernible effect on participant judgments: participants of all ages again rejected the large crowd in favor of the small group discussion for the *Easy Reasoning* questions (Younger: M=3.08, SD=.69 $t(39) = 5.35$, $p<.0001$, Older: M=3.22, SD=.51, $t(39) = 8.97$, $p<.0001$, Adult: M=2.95, SD=.78, $t(39) = 3.64$, $p=.0008$). We also observed the

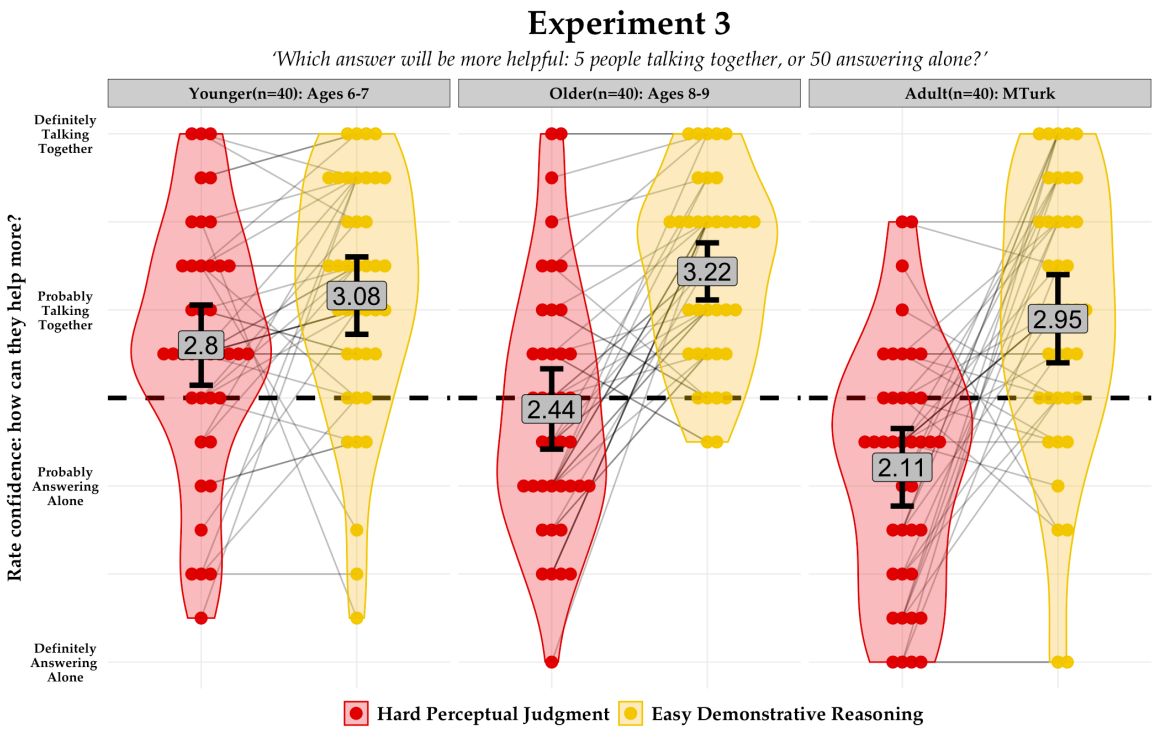


Figure 2.4. Preference for group discussion or crowdsourcing in Experiment 1; each participant's responses were averaged across four *Easy Demonstrative Reasoning* questions (yellow) and four *Hard Perceptual Discrimination* questions (red). Higher ratings indicate stronger preference for group discussion. Grey labels show means; errors bars are 95% CIs; black lines show within-subject differences for the average rating by question *Type*.

predicted developmental shift towards *Answering Alone* when reasoning was insufficient to answer the question; however, in Experiment 3 the shift occurred later than expected instead of earlier. While Adults favored *Answering Alone* and Younger children favored *Talking Together* for the *Hard Percept* questions as predicted, older children did not show the adult pattern, instead not differing from chance for *Hard Percept* (Younger: M=2.80, SD=.71, $t(117) = 2.66$ p=0.011; Older: M=2.44, SD=.71, $t(117) = -0.55$, p=0.7535; Adult: M=2.11, SD=.69, $t(117) = -3.62$, p=.0008). Moreover, while Older children and Adults distinguished between the two question *Types*, Younger children did not (Bonferroni corrected, Younger: $t(117) = -1.91$ p=.8701, Older: $t(117)$= -5.318, p<.0001, Adult: $t(117) = 5.74$, p=.0001). As in Experiments 1 and 2, the preference for group reasoning did not differ by age (all ps > .9), though Younger children showed a weaker preference than Adults for crowdsourcing *Percept* questions (Bonferroni corrected, Adult vs. Older: $t(78)$=2.156, p= *ns*; Adult vs. Younger: $t(78)$=4.516, p<0.0002; Older vs. Younger: $t(78)$=2.360, p= *ns*). Indeed, while participants of all ages were just as confident in the small group discussion for *Easy Reasoning* questions in Experiment 3 as they were for *Reasoning* questions in Experiment 2, all ages were less confident in polling a crowd of 50 for *Hard Percept* questions in Experiment 3 than for *Population Preferences* in Experiment 2 (Supplemental Materials). However, since Experiment 3 was designed contrast *Easy Reasoning* questions with *Hard Percept* questions, rather than *Hard Percept* with *Population Preferences*, these direct comparisons with Experiment 2 should be interpreted with caution: for example, the weaker preference for crowdsourcing *Hard Percept*

questions than *Population Preference* questions may be due the difference in Non-Reasoning subtype or an effect of difficulty that is specific to Non-Reasoning questions. We explore these possibilities further in the General Discussion.

2.6 General Discussion

We asked children and adults to choose between two social learning strategies: soliciting a consensus response from a small discussion group, and "crowdsourcing" many independent opinions. Though discussion can sometimes lead to groupthink, by affording individuals opportunity to correct each others' mistakes and combine insights while also reducing individual processing load, discussion can also allow small groups to outperform even their best member. In contrast, the value of crowdsourcing is fundamentally limited by the distribution of individual competence in the crowd relative to its size. The less competent individuals are on average, the larger the crowd needs to be to produce a reliably accurate estimate. Thus, when individual competence is low, crowdsourcing may be costly; when individual competence is high, the value added by crowdsourcing may have little advantage over discussion — for problems where discussion is more likely to improve accuracy than diminish it. Our results suggest that the decision to crowdsource or discuss may in part turn on learners' beliefs about the efficacy of demonstrative reasoning for a given question.

Analogously to young children's failures on false belief tasks, our results suggest that the default expectation for group judgments may be that "truth wins": though individuals may initially disagree, discussion allows groups to ultimately see the truth. As an understanding of how conscious and unconscious biases can influence people's judgments develops, learners can preempt potential biases by crowdsourcing independent judgments. Though even the youngest children in our experiments expected discussion to improve accuracy on reasoning questions, the preference for crowdsourcing non-reasoning questions underwent a developmental shift in all three experiments. Indeed, in Experiment 3, the youngest children favored discussion for both kinds of questions, suggesting that they may have failed to recognize when discussion can promote groupthink. The timing of the developmental shift is consistent with past work suggesting that between the ages of 6 and 9, children begin to use informational dependencies (Aboody et al., 2019; Sulik, Bahrami, & Roy, 2020; Magid, Yan, Siegel, Tenebaum, & Schulz, 2018; Einav, 2018) and the potential for motivational bias in individual reports (Mills & Grant, 2009; Mills & Keil, 2008) to adjudicate cases of conflicting testimony. Though recent work suggests that even preschoolers identify cases of individual bias stemming from in-group favoritism (Liberman & Shaw, 2020), unconscious biases due to herding or groupthink may be less obvious, particularly if people assume that informants are motivated to be accurate. For example, even though children as young as six *predict* that judges are more likely to independently give the same verdict when objective standards are available than when they are not (e.g., a footrace vs. a poetry contest), at age ten children are still no more likely to *diagnose* in-group favoritism as an influence on judgments in subjective contexts than objective contexts (Mills & Keil, 2005, Liberman & Shaw, 2020). In our experiments, both the reasoning and non-reasoning questions had objective answers, but only the reasoning questions afforded an objective *method of finding* those answers.

Learning to recognize this relatively subtle distinction may allow children to take advantage of the benefits of group discussion while avoiding the risks. This is not to suggest that people expect group reasoning to be infallible — merely that they expect groups to improve individual accuracy. This is consistent with recent work asking adults to predict group and individual accuracy on a classic reasoning task: while participants radically underestimated the true group advantage, they did expect groups to be more accurate than individuals (Mercier, et al., 2015). Interestingly, they also expected *dyads* to be *less* accurate than individuals. A more granular approach to intuitive beliefs about the dynamics of social influence may reveal more sophisticated intuitions: for instance, beliefs about others' conformist tendencies and the distribution of individual competence may increase confidence that "truth wins" in small groups more than in dyads.

Past work has suggested that while people dramatically underestimate crowds and overestimate their own accuracy (Mercier et al., 2020; Mannes, 2009), they defer to others more when uncertainty is high and crowds are larger (Asch, 1955). While increasing the crowd size from five to fifty had no impact on *Reasoning* questions in our experiments, the larger crowd did appear to increase crowdsourcing for *Population Preference* questions. However, while we only tested *Hard Percept* questions with a crowd of fifty, confidence in crowdsourcing was lower for *Hard Percept* questions than for *Population Preferences* (Supplemental Materials). While our design licenses no firm conclusions on this point, one reason seems evident: by definition, population preferences are whatever most individuals in a population prefer, while perceptual facts like the brightness of stars are wholly independent of individual judgments. Moreover, under the right conditions, discussing perceptual judgments with a single partner *can* improve accuracy (Sorkin, Hays, & West, 2001; Bahrami et al., 2012). Thus, participants' reduced confidence in crowdsourcing *Hard Percept* questions may have been justified. The extent to which intuitive beliefs about the benefits of discussion and crowdsourcing for different question types correspond to the empirical benefits is an open question.

Our design is limited in one important respect: the discussion group was only allowed to give a single answer, while the crowd could give multiple answers. This procedure strictly ensured that group members could not answer independently, but also entailed a unanimous consensus endorsed by a minimum of five people. Unanimous consensus can be a powerful cue: even a single dissenter can sharply reduce conformity (Asch, 1955, Whalen, Griffiths, & Buchsbaum, 2018). However, the meaning of dissent may vary across contexts and questions. In a crowd, a single "dissenter" may simply have made a mistake; but dissent-despite-discussion signals that the group has failed to convince them. When questions afford conclusive demonstrations of accuracy, failure to convince all discussants may reflect poorly on group accuracy. Conversely, in more ambiguous contexts, unanimity may suggest groupthink. For instance, in ancient Judea, crimes more likely to elicit widespread condemnation were tried by larger juries for the express purpose of reducing the odds of consensus, and unanimous convictions were thrown out on the grounds that a lack of dissent indicated a faulty process — an intuitive inference confirmed by modern statistical techniques (Gunn et al., 2016). A similar logic may underlie inferences about testimony that contradicts social alliances. For example, if Jenny says Jill is bad at soccer, even

preschoolers give Jenny's judgment more credence if Jenny and Jill are friends than if they are enemies (Liberman & Shaw, 2020). Our results suggest that even in early childhood, the absolute number of sources endorsing a belief may be less important than how those sources arrived at their beliefs. Indeed, the limited number of possible answers to the questions in Experiments 2 and 3 guaranteed that even a plurality of the 50-person crowd would considerably outnumber the 5-person group. Yet, participants' preference for discussion and crowdsourcing bore no relationship to the number of possible endorsers. Future work will compare explicit degrees of consensus in groups and crowds.

The last decade has produced an extensive literature describing how individual social learning heuristics and patterns of communication in social networks can improve or diminish collective learning (Derex & Boyd, 2016; Derex, Perreaut, & Boyd, 2018; Almaatouq, et al., 2020; Becker, Brackbill, & Centola, 2017). By focusing on population-level outcomes, much of this work has tacitly treated individuals as passive prisoners of social influence. However, the heuristics guiding social learning develop in early childhood, and recent work has shown that like other intelligent systems capable of self-organization, people are capable of "rewiring" their social networks to improve both individual and collective learning, by "following" or "unfollowing" connections depending on their accuracy (Almaatouq, 2020). Our experiments focused on two features of communication patterns that individuals can and do control in the real world, beyond *who* they choose to trust: how many people to talk to, and whether to talk with those people as a group or a crowd. Our results suggest that even in early childhood, people's judgments about how to best make use of group discussion and crowdsourcing heuristics may be consistent with the empirical advantages of each strategy. An understanding of how intuitions about social influence develop may contribute to a clearer empirical picture of how people balance the benefits of learning from collective opinion with the risks of being misled by it.

Chapter 3

Speed-accuracy tradeoffs in social cognitive inferences about individuals

This chapter is based on materials published in Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Abstract

Limits on mental speed entail speed-accuracy tradeoffs for problem-solving, but memory and perception are accurate on much faster timescales. While response times drive inference across the behavioral sciences, they may also help laypeople interpret each others' everyday behavior. We examined children's (ages 5 to 10) use of agents' response time to infer the source and quality of their knowledge. In each trial, children saw a pathfinding puzzle presented to an agent, who claimed to have solved it after either 3s or 20s. In Experiment 1 (n=135), children used agents' response speed to distinguish between memory, perception, and novel inference. In Experiment 2 (n=135), children predicted that fast responses would be inaccurate, but were less skeptical of slow agents. In Experiment 3 (n=128), children inferred task complexity from agents' speed. Our findings suggest that the simple intuition that *thinking takes time* may scaffold everyday social cognition.

3.1 Introduction

A colleague once posed a trick question to the polymath John von Neumann. von Neumann was famous for his quick mental calculations, but the solution to this particular problem could either be found by brute-force calculation, or by a rarely noticed shortcut. The question was "*Two cyclists twenty miles apart are moving towards each other at 10 mph; a fly caught between them is moving at 15 mph, from wheel to wheel and back again, until it is crushed between the wheels of the bikes. How far will the fly travel?*". The brute force approach is to sum the geometric series, adding up all of the fly's increasingly short trips between the wheels; alternatively, one could notice that the cyclists will meet in exactly one hour, when the fly will have travelled exactly fifteen miles. von Neumann answered immediately, and the colleague disappointedly replied "*oh, you've heard about the trick*". von Neumann retorted "*what trick? I simply summed the geometric series!*".

The logic of this exchange is intuitively obvious, which is remarkable given that it relies on fairly intricate series of counterfactuals and violations of expectation. Even a von Neumann could not sum an infinite series so quickly; but while both guessing and memory can be nearly instantaneous, correctly *guessing* an infinite sum is extremely improbable — so improbable that we jump to the conclusion that von Neumann must have simply recalled the trick. Discovering that von Neumann *had* summed the infinite series reveals a prodigious mental speed — and by extension, extraordinary competence with numbers. This interpretation is effortless because we intuitively understand the relative speed-accuracy tradeoffs of different cognitive processes and the significance of violating those tradeoffs.

Yet, despite the widespread use of response time as an inductive tool in the behavioral sciences, attention to laypeople's own inferences about each others' response times has been sporadic and unsystematic. Nevertheless, the few existing studies suggest that timing is a rich and flexible cue. For example, adults who respond more quickly to trivia questions are also rated as more charismatic by peers (von Hippel et al., 2016). Conversely, adults interpret longer latencies as *reluctance* in response to requests (Roberts, Francis, & Morgan, 2006), as *memory failure* in response to trivia questions (Brennan & Williams, 1995), and as *indecision* between equally desirable options in decision-making (Frydman & Krajbich, 2016, Gates et al., 2021). Moreover, in negotiation contexts, buyers' hesitation (or lack thereof) can reveal their price-point, allowing experienced sellers to adjust their selling strategy in response (Konovalov & Krajbich, 2017). Inferences like these often feel effortless despite their sophistication; yet, we are also notoriously bad at estimating the time required to complete a given task. Where do timing-based inferences come from, and how systematic are they? We suggest that even the most sophisticated inferences build on a simple intuition already present in early childhood: *thinking takes time*. On this account, seeing an agent spend more or less time on a task than expected demands explanation. Reasoning about the time costs agents incur to achieve their goals may enable the same kinds of sophisticated inferences about beliefs and desires that we make by reasoning about the costs agents incur while pursuing goals in spatial environments (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). Indeed, timing may be such a flexible cue because *faster* or *slower* responses simply index *more* or *less* thought; the question of what the agent was thinking *about* must be determined by context. Our proposal suggests a developmental approach; while young children

may expect thinking about unfamiliar problems to take more time than remembering the answer or seeing it directly, explaining why an agent took more or less time than expected in a specific context may be more difficult. We return to this point in the general discussion.

Timing-based inferences have at least three parts: (1) an observer's representation of an agent's actual response time, (2) the observer's expectations about how long a task will take to complete, and (3) a plausible explanation for any difference between the two times. While children's time perception is less precise than adults', the ability to represent and compare durations develops early: even three-year-olds judge a 3s interval between two stimuli as more similar to a 4s interval than to a 1s interval (Droit-Volet & Wearden, 2001; see Wearden, 2016, for review).

Children can also identify plausible explanations for differences in response times *between two agents*, such as competence or task difficulty. However, conflicting heuristics based on effort, speed, and outcomes often confound younger children's competence judgments until late childhood unless task difficulty is transparent (Nicholls, 1978; Leonard, Bennett-Pierre, & Gweon, 2019). For example, when experimenters *explicitly described* agents in a story as (A) finishing a puzzle quickly or slowly, (B) thinking it easy or difficult, and (C) trying hard or not, preschoolers integrated difficulty, effort, and speed (Heyman & Compton, 2006). Of course, speed, effort, and task difficulty may rarely be explicitly described in the real world. Still, more recent work suggests children can also integrate these cues spontaneously under certain conditions: when presented with videos that varied agents' speed in building block towers and the relative difficulty of their task, preschoolers recognized the tradeoffs — but only when physical cues to difficulty were unambiguous (Leonard, Bennett-Pierre, & Gweon, 2019). However, cognitively challenging tasks frequently require no physical effort at all. Nevertheless, the difficulty of cognitive processes *themselves* may be detectable for toddlers, even when the objective difficulty of the task is less certain.

Consider the case of speech disfluencies like *uh* and *um:* disfluencies occur more frequently under high cognitive load, such as might be imposed by recalling a rare word or weak memory, or by planning a complex utterance (Clark & Fox Tree, 2002; Kidd, White, & Aslin, 2011b). By 30 months, children appear to interpret speakers' speech disfluencies as resulting from processing difficulties: they predictively look at hard-to-describe or unfamiliar objects at the onset of the filled pause (Arnold, Hudson Kam, & Tanenhaus, 2007; Kidd, White, & Aslin, 2011a; Orena & White, 2015). Children's inference may reflect implicit causal reasoning: hard-to-describe objects cause processing difficulties, which in turn cause disfluencies; hence, a disfluency signals that the speaker is preparing to refer to an unfamiliar or hard-to-describe object. When given an alternative cause for a speaker's disfluency, participants' inference is blocked. For example, if a speaker frequently forgets the names of common objects, a speech disfluency may not imply that they are trying to recall a *rare* word in particular: chronically forgetful speakers may just as disfluent in producing rare words as common words. On this account, the listener's reasoning begins not from beliefs about task difficulty per se, but from recognizing the signs of effortful cognitive processing. These cues then trigger a post-hoc search for an explanation of those difficulties.

To the extent that response time signals effortful cognitive processing, it could license a variety of inferences about everyday behaviors. However, psychological processes themselves vary in the time and effort involved. For example, perceptual processes tend to be fast and automatic. Seeing someone take several seconds to respond to a question like "is this ball red or pink?" may tell us that the color they're looking at is an ambiguous case, even if we can't see it ourselves; if it turns out to be fire-engine red, we might wonder whether the person is colorblind. Memory retrieval may be slower than perception, but someone who takes tens of seconds to respond when asked their spouse's birthday might still elicit doubt or consternation, even before they produce an answer. In contrast, longer pauses are to be expected for questions that require explicit thought about complex relations: *thinking takes time*, on a scale that memory and perception only require under unusual circumstances. In short, violating the expected timescale may be conspicuous: why did the person need to *think* about what color they were seeing? Why *didn't* the person need to think about the answer to a complex calculation? Our success as individuals and as a species depends on our ability to quickly and accurately assess the knowledge, intentions, and competence of other agents; a response that is "too quick" may suggest very different inferences than a response that is "too slow".

3.2 General Method

Here, we examined the development of explicit timing-based inferences in childhood. We initially focus on children ages 5-10 because younger children may struggle to evaluate time and difficulty simultaneously for cognitive tasks (e.g., Leonard et al., 2019; Nicholls, 1978). In each experiment, participants were introduced to a pathfinding puzzle (Figure 3.1). After learning the rules, participants watched other agents play the game one by one. After either ~3 seconds or ~20 seconds, the agent signaled that he thought he knew the solution. Participants were then asked to make a judgment. In Experiment 1, participants saw a complex puzzle presented to the agent, and judged whether the agent was "figuring out the answer for the first time", or "remembering the answer from yesterday". We predicted that participants would categorize fast responses as memory, and slower responses as reasoning. Experiment 2 was identical, but participants judged whether the agent had "actually figured out" the answer or "made a mistake". We predicted that even the youngest children would expect fast responders to make mistakes, and to be more likely to make mistakes than slow responders. In Experiment 3, participants saw the agent draw a card with one of two puzzles (simple or complex), but the participants themselves could not see which; they were asked to guess which puzzle the agent was looking at, and then guess whether the agent's solution was accurate. We predicted that children would integrate response time and puzzle difficulty to infer which map the agent was looking at and whether their solution was accurate. Importantly, the puzzle difficulty and the agents' response time were never explicitly mentioned in any of the experiments; inferences based on time or difficulty were made spontaneously. All children participated through an online platform for developmental research (Sheskin & Keil, 2018). Participants came from 39 US states, were 51.7% female, 65% white, and had a median household income of $77,083, as estimated by US Census data for their reported postal code (US household median: $68,703). The pre-registrations, power analyses, data, and

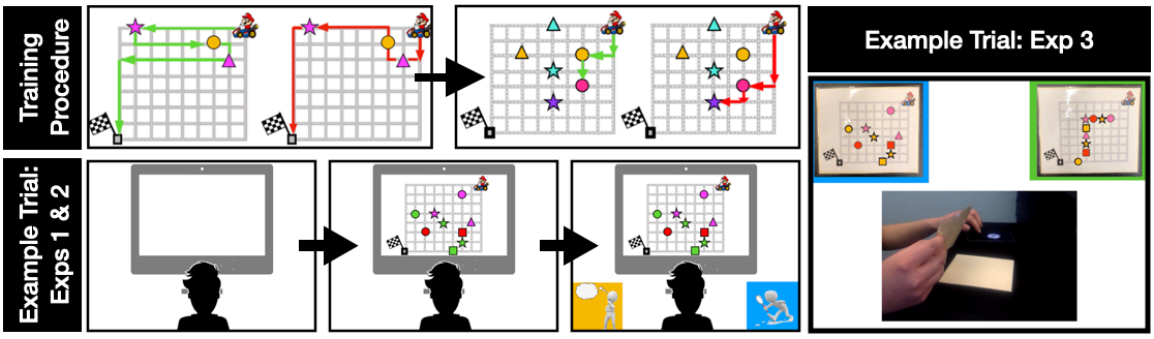


Fig 3.1. Training Procedure: Children had three chances to answer 3 of 3 training questions correctly (one example displayed): (1) "Which road is shorter, red or green? Yes, red, great job!", (2) "Which road breaks the rule? Yes, green! Great job! And *why* does it break the rule?", (3) "Let's say Mario drives on the red road instead. So first, he picks up the pink circle, and then he picks up the purple star. Which treasure should Mario pick up *next*, so that he's not breaking any rules? **Example Trials**: After a Hard map appeared on the screen (Exps 1 & 2) or the agent drew a card with either an Easy or Hard map (Exp 3), the agent signaled that they had found the shortest road that followed the rules, after either 3s or 20s.

materials for each experiment are available on the first author's OSF repository. Though our preregistered analysis plan uses standard linear regressions and ANOVAs, we also provide analogous analyses using ordinal regressions, following recent recommendations against using standard regressions for ordinal data (Liddell & Kruschke, 2018), which were brought to attention during the review process. Because both analyses produce nearly identical results in each experiment but ordinal regressions are still atypical in the field, we present our preregistered analyses in the main text and provide the ordinal regressions as a point of comparison in the supplementary materials.

3.3 Experiment 1

3.3.1 Method

Participants. We recruited 45 adults through MTurk, as well as 90 children in two age groups (45 age 5-7, M=6.5, SD=.94; 45 age 8-10, M=9.45, SD=.90; 53% girls). An additional 8 children (3 age 5, 4 age 6, and 1 age 7) and 5 adults were excluded before data collection for answering training questions incorrectly; these were replaced with new participants.

Materials. We created six grid maps with simple geometric shapes of different colors scattered across them. Each map had 8 or 9 shapes of 3 or 4 different colors. A flag in the bottom left corner of the grid marked the finish line, and a MarioKart character at the top right marked the starting line. During the test phase, participants saw the maps appear in front of a cartoon silhouette facing a computer screen. Children answered by using color-coded cartoon figures on the left and right of the screen (Figure 3.1); presentation of these answer choices was counterbalanced (Color_CB). Adults answered using a scale slider. The order of the six maps was reversed for half the participants (MapOrder_CB). Finally, four counterbalances were created to vary the order of the agents' response times (TimeOrder_CB).

Procedure. Training Phase: Participants were told that they would play a racing game with Mario, and learned the object of the game and the rules. The experimenter described the task to the children over a video-chat; the same materials were presented to adults as a prerecorded voiceover slideshow. The experimenter told the participant that Mario wanted to collect all of the treasures on a map, and take the shortest road through the map that followed a rule. The rule was that Mario could "*not pick up two treasures in a row that are the same color,* or *two in a row that are the same shape. But he has to pick up* all *of the treasures*". Participants were then required to answer four comprehension questions correctly (see Figure 3.1 caption). These were as follows. (1) Which of two example roads is shorter, (2) Which of two example roads breaks a rule, (3) Why does that road break a rule, (4) Identify an item to pick up next in an example sequence. Participants who answered each question correctly the first time proceeded to the test phase. Children's incorrect responses were gently corrected after each question, and they proceeded to the test phase only if they were able to answer all the comprehension questions correctly in two additional training rounds. Adults incorrect responses were not corrected, and adults proceeded to the test phase only if they correctly answered all the comprehension questions in the next round. After learning how to play the game, participants were told that they would watch other people playing the game, and that their job was to decide whether each person was (A) remembering the shortest road through the map from playing it the day before, or (B) figuring out the shortest road for the first time. At this point, mTurk participants also answered an attention check question in order to screen out participants who were skimming instructions.

First Task: Memory vs. Inference. For each test item (Figure 3.1), the experimenter presented a new silhouette sitting in front of a blank screen, saying "*Here's the next person. We'll show him the map, and when he thinks he knows the shortest road that follows the rules, he'll start his engine*", at which point the map appeared on the screen. After ~3s or ~20s, an engine sound played, and the experimenter said "*now he's started his engine, so he thinks he knows the shortest route through the map*", and participants decided whether the agent had been "remembering the answer from yesterday" or "figuring it out for the first time". Children first chose one of two alternatives (*remembering* or *figuring out*) and then were asked whether the agent was "*probably*" or "*definitely*" [remembering / figuring it out]; adults used a 4-point scale directly. Three maps were presented for ~3 seconds before the engine started, and three for ~20 seconds.

Second Task: Perception vs. Inference. Like memory, perceptual processes are nearly instantaneous, making direct perceptual access another potential explanation for fast responses. In a second task, the experimenter introduced two new cartoon agents, one of which was wearing opaque goggles. The experimenter specified that *neither* had played the game before (and so could not be remembering the maps), but that the silhouette with goggles "likes to cheat"; a computer in his goggles would show him the shortest road when he looked at the map, and so he would not have to figure out the answer himself. The other silhouette was described as playing fair. The experimenter presented a map to the two characters, and an engine sounded after ~3s. Participants were then asked *who* had started his engine: the one who "cheated with his special glasses and saw the answer", or the one who "played fair". We expected participants to infer that

only a cheater would have responded so quickly to a complex puzzle, but did not preregister the hypothesis for the second task.

3.3.2 Results

The three responses at each response speed were averaged to create a single score for each (Figure 3.2a). A repeated measures ANOVA revealed a significant effect of Speed ($F(1,132)=202.09$, $p<.0001$, $\eta^2 = .605$) and an AgeGroup*Speed interaction ($F(2,132)=17.51$, $p<.0001$, $\eta^2 = .210$), but no effect of AgeGroup ($F(2,132)=0.425$, $p=.65$, $\eta^2 = .006$). As predicted, all age groups categorized the fast response as memory ($M_{Young}=2.13$, $t(44) = -3.23$, $p=.002$, $M_{Old}=1.80$, $t(44) = -7.33$, $p<.0001$, $M_{Adult}=1.47$, $t(44) = -13.00$, $p<.0001$), and categorized the slow response as inference ($M_{Young}=2.75$, $t(44) = 2.63$, $p=.012$, $M_{Old}=3.21$, $t(44) = 7.23$, $p<.0001$, $M_{Adult}=3.45$, $t(44) = 10.58$, $p<.0001$). All age groups also identified the fast responder as having cheated (Figure 3.2b), including 80% of 6 year olds and 86.7% of 7 year olds, suggesting that even the youngest children recognized that three seconds is an impossibly fast latency to solve the puzzle for the first time, but is easily explained by having direct perceptual access ($M_{Young}=71.1\%$, binomial $p=.003$, $M_{Old}=84.4\%$, binomial $p<.0001$, binomial $M_{Adult}=91.1\%$, $p<.0001$).

Next, we explored whether children would rate the fast response as memory at younger ages than they rated slow responses as inference, but our prediction here was not supported. For each subject's average rating for fast and slow responses, we calculated the deviation from chance responding. We then regressed these values on age in years, using contrast coding to compare

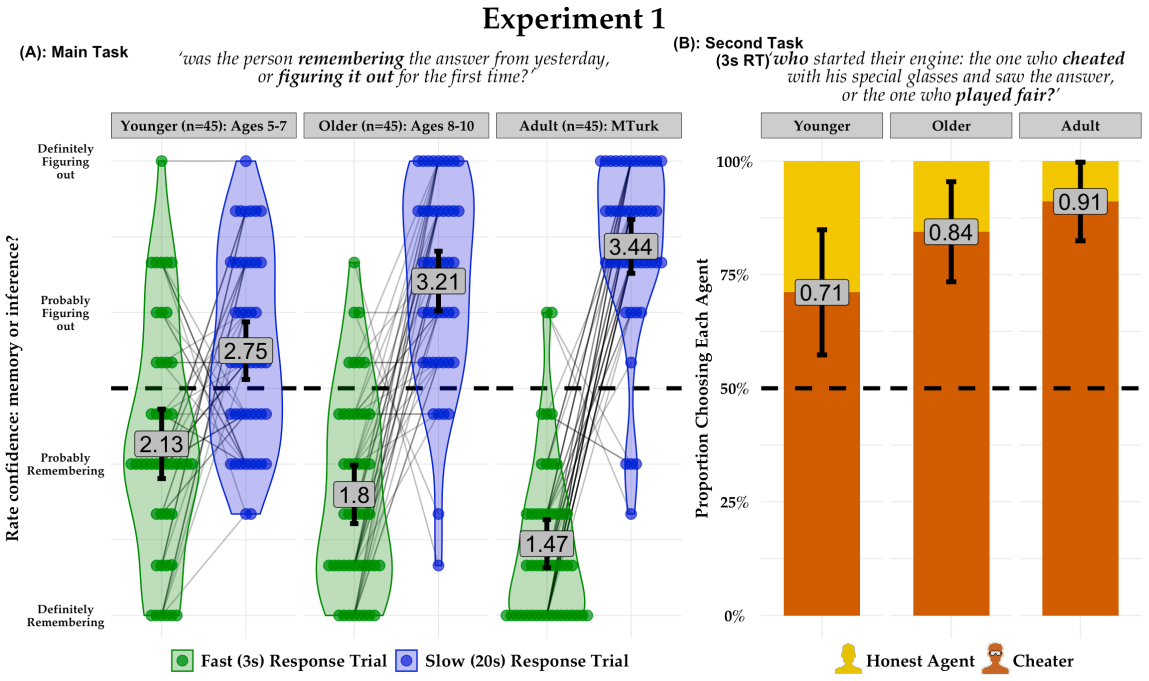


Figure 3.2a-b. (a) Results from Task 1 (Memory vs. Inference). Violin plots and means with 95% CIs for Fast and Slow trials in Experiment 1. Each dot is the average of each participant's 3 fast trials (Green) or 3 slow trials (Blue). Participants rated each agent on a 4-point scale. **(a) Results from Task 2 (Perception vs. Inference).** Bar plots with means and 95% CIs from the "cheater" task.

each level to chance. The 7, 9, and 10 year olds were more likely to rate the *fast* responses as *memory* (β_7 = .66, p<.001, β_9 = .88, p<.001, β_{10} = .95, p<.001), but the 5, 6, and 8 year olds did not differ from chance (β_5 = .21, p=.226, β_6 = .23, p=.180, β_8 = .28, p=.111). The 7, 8, 9, and 10 year olds were more likely to rate the *slow* responses as *inference* (β_7 = .54, p<.001, β_8 = .37, p=.020, β_9 = .70, p<.001, β_{10} = 1.06, p<.001), but the 5 and 6 year olds did not differ from chance (β_5 = -0.08, p=.618, β_6 = .28, p=.075).

These results evince an early-developing commonsense intuition that "thinking takes time", while perception and memory — even memory for a solution to a complex problem — are expected to be much faster. In Experiments 2 and 3, we ask whether children can use this intuition to predict the accuracy of an agents' response, and whether they modulate their judgments according to the difficulty of the problem.

3.4 Experiment 2

3.4.1 Method

Participants. We recruited 45 adults through mTurk, as well as 90 children in two age groups (45 age 5-7, M=6.79, SD=.82; 45 age 8-10, M=9.84, SD=.82; 51% girls). An additional twelve children and one adult were screened out and replaced before data collection for failing the training (6), losing internet connection (2), fussing out (2), parent interference (2), and colorblindness (1).

Procedure. First Task: Speed-Accuracy Tradeoffs. We made one change to our materials from Experiment 1. Agents were now described as playing the game for the first time, and participants were asked to guess whether each agent had "*actually* figured out the shortest road, or if they *made a mistake*", again using a 4-point confidence scale. Second Task: Speed & Competence. To compare our results to past work on children's timing-based inference, one trial at the end of the experiment asked participants to judge the relative competence of two agents who each accurately solved the same puzzle after either 3s or 20s. Because this task was included simply to compare our results with past work, the full method and results are described in the Supplemental Materials. In brief, younger children and adults were equally likely to judge the fast and slow agent as "better at this game", but older children believed the fast agent was better.

3.4.2 Results

Results are shown in Figure 3.3. The three responses at each response speed were averaged to create a single score for each. A repeated measures ANOVA revealed a significant effect of Speed ($F(1,111)$=72.16, p<.0001, η^2 = .394) and an AgeGroup*Speed interaction ($F(2,111)$=10.69, p<.0001, η^2 = .162), but no effect of AgeGroup ($F(2,111)$=1.50, p=.23, η^2 = .026). There were two unexpected order of presentation effects, both suggesting that the predicted effect was larger for one counterbalance than the others; however these effects were smaller than the main effect of Speed, and subsequent analyses suggested that they could not explain the focal findings (TimeOrder_CB: $F(3,111)$=5.27, p=.002, η^2 = .125; MapOrder_CB*Speed: $F(1,111)$=3.97, p=.049, η^2 = .034). Post hoc comparisons of the Age*Speed interaction revealed that while the adults and older children distinguished between 3s and 20s responses as predicted, the difference for younger

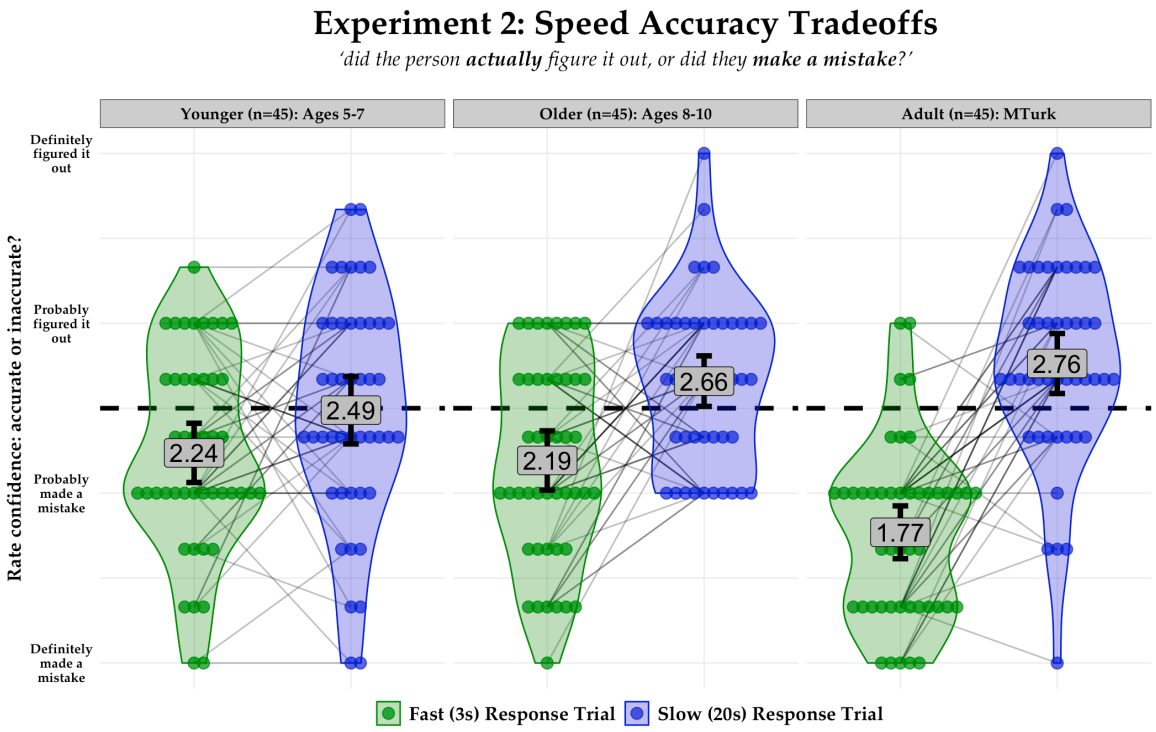


Figure 3.3. Violin plots and means with 95% CIs for Fast and Slow trials in Experiment 2. Each dot is the average of each participant's 3 fast trials (Green) or 3 slow trials (Blue).

children was not significant (Young: M_{Fast}=2.24, M_{Slow}=2.49 $t(111)$ = -2.18, p=.473; Older: M_{Fast}=2.19, M_{Slow}=2.66, $t(111)$=-4.02, p=.0016; Adult: M_{Fast}=1.77, M_{Slow}=2.76, $t(111)$ = -8.49, p<.0001, bonferroni corrected). However, all age groups predicted that fast responses were likely to be wrong (M_{Young}=2.24, $t(44)$=-3.04, p=.004, M_{Old}=2.19, $t(44)$ = -3.55, p<.001, M_{Adult}=1.77, $t(44)$ = -9.47, p<.0001). While adults and older rated the slow responses as likely to be right, younger children did not differ from chance (M_{Young}=2.49, $t(44)$=-0.12, p=.907, M_{Old}=2.66, $t(44)$=2.16, p=.036, M_{Adult}=2.76, $t(44)$ = 3.00, p=.004). Children may have been right to be skeptical of accuracy on the slow trials: given the computational complexity of these problems, even 20s is too fast to solve them except by luck. Indeed, given past work suggesting that even older children can be unreasonably credulous towards confident speakers (Kominsky, Langthorne, & Keil, 2015), even a skepticism on the fast trials that emerges around age 6 or 7 may be precocious. However, since the difference between the fast and slow trials was not significant in the younger age group, interpretations of the youngest children's responses as reflecting skepticism of the agents' accuracy should be taken with a grain of salt. In Experiment 3, we examine the possibility of an early-but-nuanced skepticism more closely, asking whether children's judgments integrate both response time and task difficulty.

3.5 Experiment 3

3.5.1 Method

If observers infer that "*more time = more effort*", then they may infer that agents who spend more time are solving more complex problems than agents who spend less time. We predicted that children would infer that (A) fast agents were more likely to be looking at easy maps than hard maps, (B) fast agents were more likely than the slow agents to be looking the easy maps, and (C) fast agents' solutions were correct for the easy maps but incorrect for the hard maps. Because Experiments 1 and 2 suggested that timing-based inferences appear to emerge around age 6 or 7, we focused on ages 6-8 in Experiment 3.

Participants. We recruited 32 adults through mTurk, as well as 96 children (32 age 6, M=6.46, SD=.31; 32 age 7 M=7.55, SD=.31; 32 age 8, M=8.59, SD=.30; 48 girls). An additional 13 children were screened out and replaced before data collection for failing the training (7), technical difficulties preventing videos from playing (5), and fussing out (1).

Materials. We generated a set of *Easy* puzzles by rearranging the treasures on the complex puzzles from Experiments 1 and 2 into a row of alternating colors and shapes, so that the shortest route passed directly through them. This produced pairs of maps which were identical in the number and kind of treasures, but were *Easy* or *Hard* to solve. In each trial, participants saw an agent draw a card with one of the two maps; when the agent rang a bell to signal that they were ready, children guessed which map was on the card. As in previous experiments, neither time nor the difficulty of the maps was ever explicitly mentioned.

Procedure. The training phase was the same as in Experiments 1 and 2. First Task: Which Map? Agents were described as playing the game for the first time. The experimenter explained that each person would take a card with a map on it, and ring a bell when they thought they knew the shortest road. On each trial, the experimenter showed the participant a slide with an Easy map a Hard map, and a video embedded between them, and reminded the participant of the task: "this person might get the card with this map [points at simple map] or they might get the card with this map [points at hard map]. Your job is to guess which map was on their card". After the agent rang the bell, the experimenter said "*They rang the bell, so that means that they think they've figured out the shortest road through the map. But which map was on the card they got?*". Children first chose one of the two alternatives and then were asked whether the agent was "*probably*" or "*definitely*" looking at the map; adults used a 4-point scale directly. In two Fast trials, the agent rang the bell after 3s, and in two Slow trials the agent rang the bell after 20s. The order of the trials and the color of the answers was counterbalanced.

Second Task: Difficulty & Accuracy. After the main task, participants completed two additional Fast trials. In one video, both cards had Hard maps. In the other, both had Easy maps. Participants were first asked which map the agent was looking at, but then also guessed whether the agent's solution was correct or not, on a 4-point scale. The order of the Easy and Hard trials was counterbalanced.

3.5.2 Results

The two ratings at each response speed were averaged to create a single score for each. The primary question of interest (Figure 3.4) was whether children would infer that the *Fast* agent was looking at the *Easy* puzzle. To test this, we centered children's average ratings on the Fast

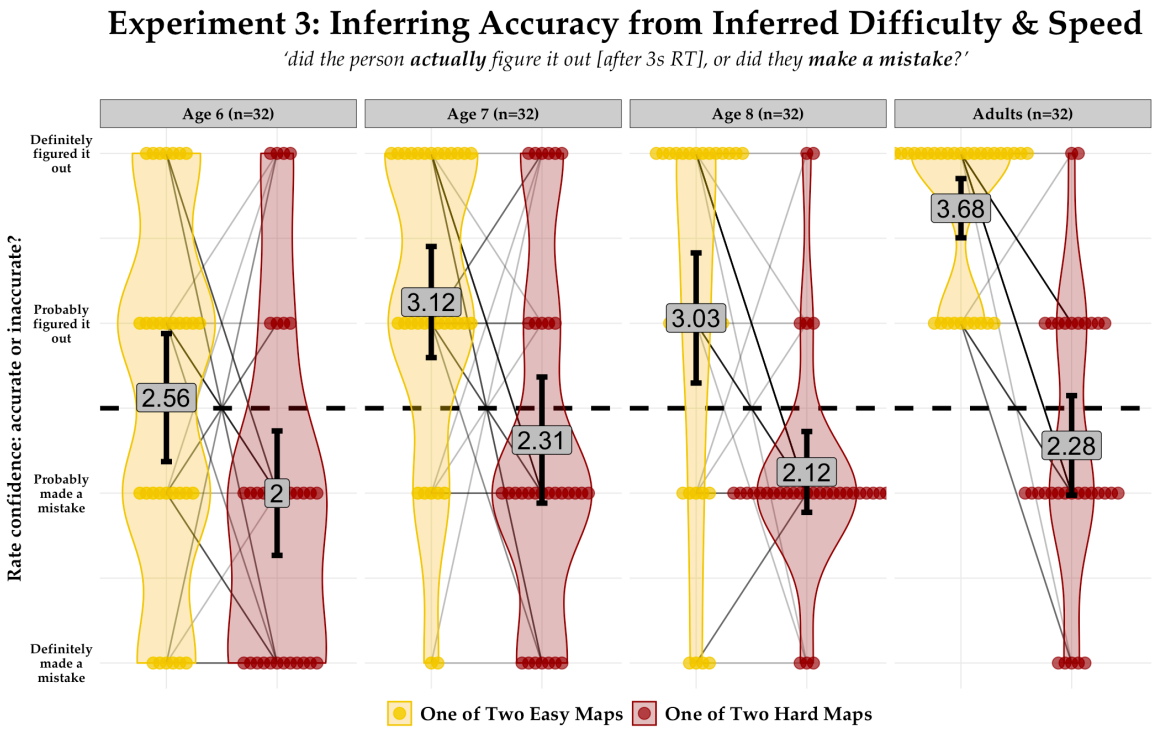


Fig 3.5. Violin plots and means with 95% CIs for two Fast trials in Experiment 3 in which participants first inferred which of two Hard or two Easy puzzles the agent was solving, and then predicted the agent's accuracy on a 4-point scale. Though participants were equally likely to guess either puzzle when the difficulty was equalized, they predicted that the fast agent's solution was correct for Easy puzzles and incorrect for Hard puzzles. Note that Easy and Difficult were left implicit; the complexity of the map was never explicitly mentioned.

trials on chance performance (2.5 on a 4-point scale), and centered children's age on the average age of the sample. This makes the intercept of the regression equivalent to a t.test for the whole sample, but allows us to simultaneously check for age effects, using age as a continuous variable. As predicted, children were more likely to infer that the Fast agent must have been looking at the *Easy* puzzle than the *Hard* puzzle (β_{Int} = -0.68, SE = .076, p<.0001); the age effect was also significant, though smaller (β_{Int} = -0.21, SE = .093, p<.026). To examine the developmental pattern more closely, we also conducted one-sample t.tests comparing each age to chance separately; all ages were significantly more likely to infer that the *Fast* agent was looking at the *Easy* puzzle than the *Hard* puzzle (M_{Age6} = -0.53, t(31)=-3.38, p=.002; M_{Age7} = -0.56, t(31)=-4.13, p=.00026; M_{Age8} = -0.95 t(31)=-10.2, p<.0001). The effect was similar for adults (M_{Age8} = -1.38 t(31)=-25.0, p<.0001). Next, we compared children's inferences for *Fast* and *Slow* agents, using *AgeYears* and *Speed* as predictors. A repeated measures ANOVA revealed a significant effect of *Speed* (F(1,93)=28.11, p<.0001, η^2 = .232) but no effect of *AgeYears* or *AgeYears*Speed* interaction (*AgeYears*: F(2,93)=1.45, p=.24, η^2 = .03; *AgeYears*Speed*: F(2,93)=1.31, p=.27, η^2 = .027). Paired-sample t.tests revealed that the effect was similar for all age groups individually, including adults (*Age6*: M_{Fast}=1.97, M_{Slow}=2.41, t(31) = 14.1, p<.0001; *Age7*: M_{Fast}=1.94, M_{Slow}=2.58, t(31) = 14.2, p<.0001; *Age8*: M_{Fast}=1.55, M_{Slow}=2.48, t(31) = 16.3, p<.0001; *Adults*: M_{Fast}=1.12, M_{Slow}=3.88, t(31) = 70.5, p<.0001).

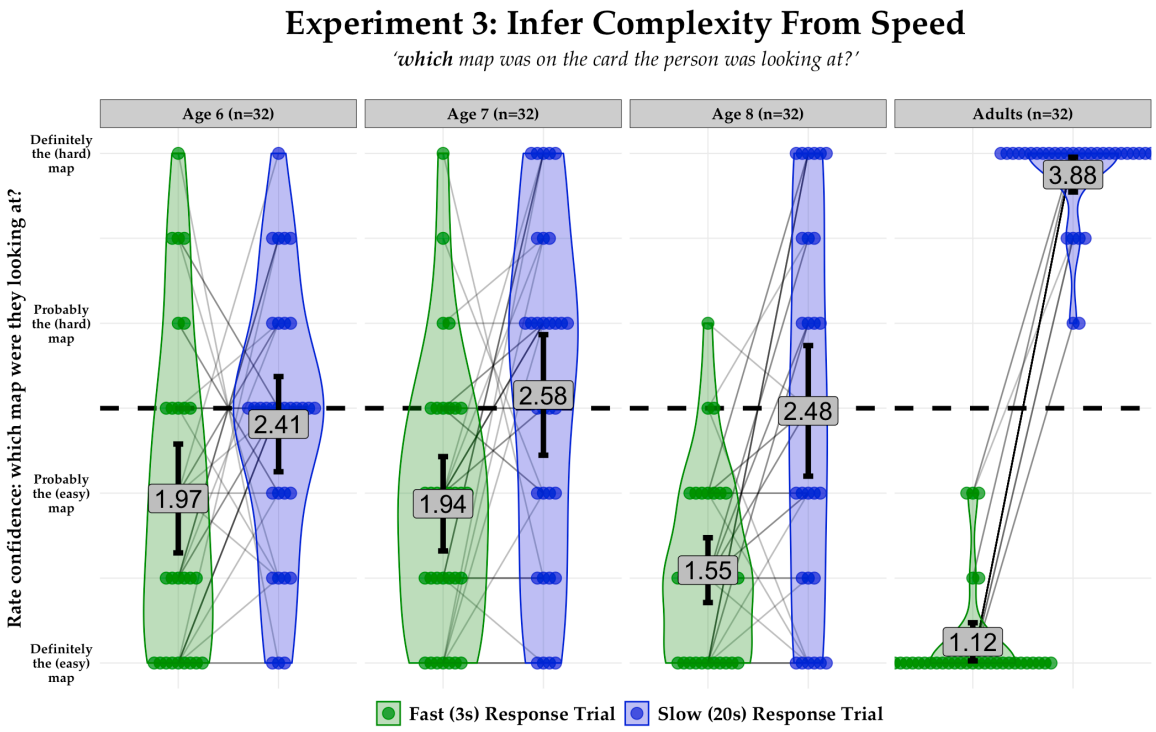


Fig 3.4. Violin plots and means with 95% CIs for Fast and Slow trials in Experiment 3. Each dot is the average of each participant's 2 fast trials (Green) or 2 slow trials (Blue). After watching an agent draw a card, look privately, and ring the bell to indicate when they think they know the shortest route, participants were asked which map the agent was looking at. Note that Easy and Difficult were left implicit; the complexity of the map was never explicitly mentioned.

Children's judgments thus appear to integrate both the complexity of the puzzle and the agent's response speed. This conclusion is further corroborated by the results of the second task (Figure 3.5), in which a *Fast* agent drew one of two *Easy* puzzles or one of two *Hard* puzzles: deprived of task difficulty as a cue, participants were no more likely to infer that the agent had drawn one than the other, in any age group (all *p's*=n.s.; see Supplemental Materials). However, all ages were more likely to say that the agent had solved the *Easy* puzzle than the *Hard* puzzle (Difference scores: $M_{Age6} = 0.56$, 95 CI: 0.03–1.09; $M_{Age7} = 0.81$, 95 CI: 0.25–1.31; $M_{Age8} = 0.91$, 95 CI: 0.50–1.31; $M_{Adult} = 1.35$, 95 CI: 1.06–1.61), suggesting that they recognized the relative difficulty of the two puzzles. Estimations of absolute difficulty were less clear. Adults and children ages 7 and 8, but not age 6, believed that the agent's solution was correct for the *Easy* puzzle ($M_{Age6} = 2.56$, 95 CI: 2.22–2.91; $M_{Age7} = 3.12$, 95 CI: 2.81–3.41; $M_{Age8} = 3.03$, 95 CI: 2.66–3.38; $M_{Adult} = 3.68$, 95 CI: 3.52–3.84), while children ages 6 and 8, but not adults or 7-year-olds, believed that the agent's solution was incorrect for the *Hard* puzzle ($M_{Age6} = 2.00$, 95 CI: 1.66–2.38; $M_{Age7} = 2.31$, 95 CI: 1.97–2.69; $M_{Age8} = 2.12$, 95 CI: 1.91–2.38; $M_{Adult} = 2.28$, 95 CI: 2.00–2.56).

3.6 General Discussion

Response time has been a powerful inductive tool in the behavioral sciences. It has been used to infer preference strength (Konovalov & Krajbich, 2019), intelligence (Salthouse, 1996), the strength of memory traces (Singer & Tiede, 2008), and of course, diligence in online surveys.

Response times have even been argued to impose bottom-up constraints on models of perception, by comparing the maximum transmission speed of a single neuron with the time typically sufficient for basic perceptual tasks (Feldman & Ballard, 1982). Less attention has been paid to how laypeople themselves interpret response times.

Our experiments provide evidence that from an early age, the commonsense intuition that "*thinking takes time*" may help us interpret everyday behaviors. Indeed, Experiment 3 suggests that children spontaneously integrate task difficulty to estimate *how much* time a task should take: all ages expected slower responses to harder problems than easier problems. These estimates may help children decide *how fast is too fast* for an agent solving a novel problem. Children appeared to recognize speed-accuracy tradeoffs and modulate their accuracy judgments according to task difficulty (Experiments 2 and 3). Moreover, when confronted with quick responses to hard problems, children believed that the agent must have recalled the answer from memory or seen it directly (Experiment 1). Given participants' propensity to explain away fast responses as inaccurate, memory-based, or simple, it may seem inconsistent that only 8-10-year-olds inferred that agents who quickly solved complex novel problems were more competent than slower agents (Experiment 2). However, given the complexity of the hard puzzles, adults' judgments may simply reflect the more sophisticated judgment that the fast agent had only "solved" the puzzle by a lucky guess.

The possibility of 'lucky guesses' illustrates an interesting contrast between reasoning about agents' allocation of time and reasoning about their navigation of space: costs measured in time may be more malleable than costs in distance, making reasoning about the utility of agents' actions on the basis of time more challenging — but also potentially more informative about the agent themselves. An agent that prefers a reward that is spatially distal over one that is spatially proximal *must* pay a higher cost to obtain the more valuable reward *every* time they do so; in this sense, space imposes a fixed cost to any physical action (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015). For instance, if the toddler in the dining room prefers the cherries in the kitchen to the dinner on the table, the *distance* to the kitchen is fixed both for them and their parent: no walking, no cherries. The immutability of spatial costs may make them particularly useful for analyzing rational action even in early infancy (Gergely & Csibra, 2003; Liu et al., 2017). In contrast, while it may take longer to *count* a bowl of 35 cherries than a bowl of 30 cherries, a lucky guess could get them the 35 cherries without having to count; moreover, developmental changes in counting skill and precision in approximate number estimation may lead to different expected values of each strategy for a toddler and their parent (Halberda & Feigenson, 2008; Baer & Odic, 2019). Analogously, if an expert gives a quick estimate instead of a time-consuming calculation, we might infer that an 'educated guess' is sufficiently precise; but a novice who gives a quick estimate may not even understand the parameters of the question. In other words, while guessing and reasoning have their characteristic time signatures, successfully reasoning about time-costs may require us to consider context-specific factors like the complexity of the problem, the methods available to solve it, and potentially the competence of the agent. Thus, even with hard

constraints on mental speed, cost-based reasoning about time may require more sophisticated inferences than cost-based reasoning about space.

Timing's sensitivity to context may also help explain why competence judgments show a protracted developmental trajectory in the existing literature (Nicholls, 1978; Stipek & Mac Iver, 1989; Heyman & Compton, 2006): integrating multiple cues can be challenging for children. Neither speed, effort, nor accuracy alone signals competence: a competent agent must *outperform* the expected speed-accuracy tradeoffs *because of* their abilities. When simple heuristics like "more effort = better outcomes", "better outcomes = more competent", and "faster = better" imply conflicting competence judgments, children may find it difficult to weigh the relative importance of each dimension. Accounting for differences in motivation and attention adds to the challenge: conscientiousness and distraction can both increase response time, just as genius and haste can decrease it. However, recent computational work has suggested that by adulthood, people can distinguish distraction from focused thought by integrating the response time and complexity of the most likely topic of focus (Berke & Jara-Ettinger, 2021).

Children's inferences in our experiments were less sophisticated than competence judgments, but also more general: thinking takes more time than remembering or seeing, more complex problems require more thinking, and some problems are impossible to solve immediately. However, while our aim was not to establish the earliest age at which children can reason about response time, task demands still limit our conclusions about younger children's abilities. Though the rules to the puzzle game we used were simple enough that most children had no trouble learning them, the game was novel to children, and the procedure provided little reinforcement of the rules after the training. While our participants displayed precocious skepticism and sensitivity to task difficulty, novelty and low incentives may have hindered younger children's performance. If children learn to simulate others' mental processes through experience of their own mental processes in similar contexts, our study may have underestimated their capacities simply by giving them little experience solving the maps themselves (Sommerville, Woodward, & Needham, 2005; Meltzoff & Brooks, 2008; Kano et al., 2019). While the task was also novel for older children and adults, they may have also found it easier to simulate solving the maps for themselves before answering. Future work could test children's performance in a more familiar context, or compare their performance with and without additional practice solving similar puzzles.

Future work could explore the impact of inferences about time on children's learning strategies. Some of these inferences may come from monitoring their own response time. For example, children increasingly modulate the time spent on easy versus difficult items in the Raven's Progressive Matrices battery with age, and the degree of modulation is strongly correlated with performance (Perret & Dauvier, 2018). By adulthood, people's problem-solving strategies not only weigh time costs in the hundreds of milliseconds, but may integrate both cognitive and physical costs (Gray, Sims, Fu & Schoelles, 2006; Feghhi & Rosenbaum, 2019). Choices between different strategies can be thought of in terms of opportunity costs: in addition to costs and benefits of each strategy individually, an individual who uses one strategy foregoes the opportunity to benefit from the other strategy (Boreau, Sokol-Hessner, & Daw, 2015). Thus,

an understanding of how children learn to allocate time effectively may need to consider the both effectiveness of the problem-solving strategies available to them and the cognitive and physical tradeoffs between those strategies. For instance, in contexts that provide immediate accuracy feedback and little penalty for mistakes, it may be more rational to learn by trial-and-error than attempting to solve a problem through thinking alone. Future work could explore how children allocate time when the costs of error are high or low. Future work could also explore the use of response time in combination with other common social learning strategies. For instance, novices may be generally slow; but delays from experts may indicate a complex problem, a valuable solution, or a gap in the field's knowledge. Thus, experts' time allocation in particular may help learners estimate the value of persistence, either generally or for a specific task or problem-solving method. Indeed, children persist longer at physical tasks after observing adults spend more time and effort, but only if the adult's persistence paid off (Leonard, Lee, & Schulz, 2017; Leonard, Garcia, & Schulz, 2019).

Even in early childhood, the assumption that agents pursue goals efficiently by minimizing expected costs while maximizing expected rewards helps us reason about others' preferences, knowledge, and beliefs (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Gergely & Csibra, 2003). Much of this work has focused on the costs imposed by navigating complex spatial environments. As a fundamental constraint on every cognitive process and social interaction, time imposes costs that are even more ubiquitous, but may be more challenging to evaluate because of their sensitivity to context. Nevertheless, our results suggest that by age 6, the commonsense intuition that 'thinking takes time' — more time than perception and memory — may allow us to infer *how* another agent knows something as well as the *quality* and *complexity* of their knowledge. As children learn to integrate contextual information such as agents' expertise and the difficulty of a problem, this simple intuition could scaffold more sophisticated reasoning about agents' knowledge, intentions, and reliability.

Chapter 4

Speed-accuracy tradeoffs in social cognitive inferences about groups

This chapter is based on materials taken from Richardson, E., Hok, H., Shaw, A., & Keil, F. C. (*in prep*). Herding cats: Children's intuitive theories of persuasion predict slower collective decisions in larger and more diverse groups, but disregard factional power. submitted to *Proceedings of the Cognitive Science Society*

Abstract

Collaboration can make collective judgments more accurate than individual judgments, but it also comes with costs in time, effort, and social cohesion. But how do we estimate these costs? In two experiments, we introduce children and adults to two teams in which the teammates disagree about the optimal solution to a novel problem, and ask which team would need more time to reach a consensus decision. We find that all ages expect slower decisions from teams with more people or factions, and expect the number of factions to matter more than the number of people. But only adults expect decisions initially endorsed by a stronger faction to be faster than those endorsed by a weaker faction. Results are discussed in context of children's reasoning about social power, and models of time-rational collective decision-making.

4.1 Introduction

Reaching consensus can feel akin to herding cats: time-consuming and sometimes hopeless. But the struggle's not unique to committees of colicky faculty or poorly managed advisory panels. Differences of opinion are inevitable in groups, and time spent debating those differences adds up. Since people may agree on *what* to do without agreeing on *why*, discussions can easily involve more opinions than people, even in groups debating a yes-no decision about a single option. While some of those debates are sure to be more substantive than others, the clock ticks just as quickly for groups quibbling over minutiae as groups deliberating about substantive issues. And since one person's molehill may be another's mountain, dissent could continue to undermine consensus indefinitely. But it doesn't. We're not cats, after all; humans excel at collaboration and coordination. By adulthood, it seems commonsensical that collaborators need to weigh the costs of debate against the benefits. In some cases, getting consensus on your side may simply be too unlikely or too time-consuming to make a difference of opinion worth debating.

Clearly, the social dynamics that drive collective decision making are complex. But reasoning about how they contribute to decision speed doesn't seem to require much effort. For instance, it seems commonsensical that large groups will need more time than small groups to make decisions, or that groups in which a strong initial consensus can pressure dissenters to concede will make decisions more quickly than groups facing multipolar negotiations with no initial consensus at all. We suggest that these inferences feel effortless because they are generated by an intuitive theory (or suite of them) which inputs our beliefs about the constraints on a group decision and outputs systematic inferences about the ways we can influence the group's opinion dynamics — including outcomes, but also costs in time, effort, and social cohesion. Intuitive theories may begin as little more than a few salient cues and some beliefs about their causal connections (Keil, 2011; Mahr & Csibra, 2022), and they don't need to be particularly accurate or precise. They simply need to allow us to navigate a conceptual domain in everyday life, and be flexible enough to accommodate conceptual change and development.

Here, we provide evidence of systematic inferences about group decision speed in children and adults. We suggest these inferences may emerge from a few causally-connected intuitions about how group decision-making works: (1) expressing an opinion takes time, (2) debating differences takes even more, and while (3) not every difference of opinion is worth debating, a team's size and structure can make the cost-benefit tradeoffs of debate different for different teammates.

To illustrate how these intuitions generate predictions about decision speed, consider a robotics team deliberating over seven kinds of propeller for a drone (Figure 4.1). Discussion takes time, but any teammate can *concede* whenever they want — either because they've been convinced or because they simply don't think it's worth arguing further. However, one person's unilateral concession is only guaranteed to *save* time in Panel 1, where the debate will end as soon as either teammate concedes (assuming the other doesn't). By contrast, out of the five teammates in Panel 2, only one person can end the debate unilaterally by conceding: after all, even if one of the other four conceded, their former allies could continue to argue. And in every other Panel, no single person can unilaterally end the debate: the teammates *have to* spend time coordinating within and

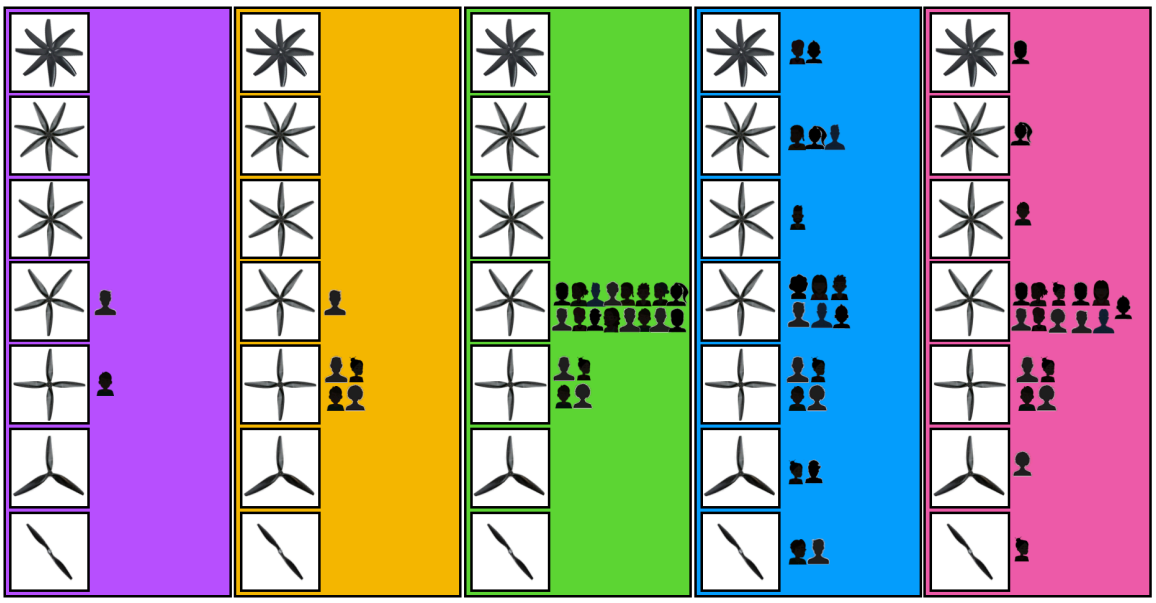


Figure 4.1. Five different robotics teams divided into two or more factions. Since they disagree, they need to talk together to make a decision about which propeller to use. Panels 2-3 have the same proportional distributions; teams in Panels 3-5 are the same size, but the faction endorsing the 4-blade varies in factional power across panels while representing the same proportion of teammates

across factions in order to reach *any* consensus, regardless of whether they're arguing for their own propeller or simply trying to find an expedient option.

In short, the harder it is for teammates to predict each others' behavior, the more coordination will be needed to make a decision; and the more coordination required, the slower the team will be. These intuitions aren't simply illusions. Agent-based models demonstrate that increasing a group's size or preference diversity leads to slower decisions; and while lower decision thresholds (e.g., plurality or majority instead of supermajority or unanimity) can speed up decisions, they can also give stubborn minorities leverage over moderate majorities if swing voters are impatient (Albrecht, Anderson, & Vroman, 2010; Chan, Lizzeri, Suen, & Yariv, 2018). Our story is simply that if our intuitive theories make reasoning about these dynamics relatively effortless, we may be able to make more rational use of our time and effort in collective action. But the constraints on group decisions our intuitive theories are most sensitive to may change from early development to adulthood. We'll return to this point in a moment.

We predict that adults will expect slower decisions from teams with more people or factions. But we also predict that they'll also expect quicker decisions from teams in which consensus is already strong at the outset than from teams in which power is initially more equally distributed between factions. Why? Because consensus is not just an outcome; it's also an epistemic and normative influence on people's responses to disagreement (Morgan & Laland, 2012; Kameda, Toyokawa, & Tindale, 2022). For instance, adults defer more to polls showing a 16v4 majority than either a proportionally weaker 11v9 majority or numerically weaker 4v1 majority (Mannes, 2009). Being mutually aware that a team is converging on a consensus decision may put dissenters under increasing pressure to concede even if they disagree, and make the strongest

faction increasingly difficult to fracture. Importantly, factional power is more than just degree of consensus. In teams with multiple factions, one faction's power over another may depend as much on their dynamics with a third as their own size or proportion. In other words, factional power is a matter of "party discipline" (i.e., how strictly individuals subordinate their idiosyncrasies to the interests of their factions) as well as conformist tendencies (i.e., deference to consensus). However, conceptual development in early childhood often produces qualitative changes in our intuitive theories. Here, we examine 6-to-9-year-olds inferences about how the number of people, factions, and consensus strength affect decision speed.

We predict that young children, like adults, will expect slower decisions from teams with more people or factions. Why? First, talk takes time, and reasoning about the relationship between time, effort, and task difficulty emerges early. Even four- and five- year-olds expect agents to take longer to complete more difficult physical tasks (Leonard, Bennet-Pierre, & Gweon, 2019); and by age six, children also begin to expect agents to take longer to solve more complex reasoning problems, unless the agent has seen the solution before (Richardson & Keil, 2022). So, children may infer that having more factions or people on a team leads to slower decisions because it makes coordination into a more complex or effortful task. Second, children may also be able to infer how much talk goes into resolving disagreements by drawing on their own experience of collaborative reasoning. Three-year-olds explicitly dispute statements they believe to be false (Köymen & Tomasello, 2018), but how much of their reasoning they verbalize depends on what they expect their collaborators to know already (Köymen, Mammen, & Tomasello, 2016). And while preschoolers do evaluate each others' reasoning, they only begin to engage in meta-talk comparing higher-order evidence such as their relative confidence or their informants' reliability between the ages of five and seven (Köymen & Tomasello, 2018). Along with believing that difficult tasks take longer and that coordination gets harder in larger groups, being inefficient collaborators could make children especially sensitive to how increasing the number of people or factions on a teams can slow down collective decisions.

However, reasoning about how consensus strength impacts decision speed may be more challenging for children. Why? First, at least one mechanism that allows adults to speed up group decisions seems to be less reliable in children: while preschoolers conform to majority opinion in both informational and normative contexts, *stronger* deference to proportionally *larger* majorities only emerges around age six or seven, even with only two factions to consider (Morgan, Laland, & Harris, 2015). That is, preschoolers are no more deferential to a 9v1 majority than a 6v4 majority — and they are selective about when they defer to majorities to begin with (Burdett et al., 2016; Haun, van Leeuwen, & Edelson, 2013, Pham & Buchsbaum, 2020). And children don't simply become more deferential; they also become more selective about deferring. Though seven year olds are more likely to defer when uncertainty is high, they are also more likely to point out when they think the emperor is clearly naked (Morgan et al, 2015). Second, strategic deference in group contexts is rarely just a matter of votes; it often depends on how we evaluate each others' approximate explanations of matters we only partially understand to begin with (Keil, 2006). Children are much less skilled than adults in adjudicating conflicting explanations, and often strikingly overconfident in their own knowledge (Kloo, Rohwer, & Perner, 2017; Mills & Keil,

2004). Taken together, these findings suggest that (1) disputes over idiosyncratic and fundamental differences may not be as strictly triaged or efficiently resolved in groups of children as in groups of adults, and that (2) at least one mechanism that speeds up decisions in adults — stronger epistemic deference to stronger consensus, particularly without argument — may be less reliable in children. Thus, while children may expect slower decisions from teams with more factions or more people, they may not expect consensus strength to increase decision speed.

To be clear, however, the claim is not that children fail to recognize differences in consensus strength per se. Even preschoolers can accurately represent and compare small differences in numerical sets (Halberda & Feigensen, 2008). Moreover, we think it's clear that children can make some inferences about power from relative group size (Pun, Birch, & Baron, 2016; Heck, Bas, & Kinzler 2021). For instance, by 6-9 months, infants expect an agent with only one ally to make way for an agent with two allies, even if the one ally is physically larger than the two allies put together (Pun, Birch, & Baron, 2016). And preschoolers infer that even though larger groups are more likely to "get the stuff", smaller groups are more likely to "be in charge" — suggesting that children not only recognize the strength in numbers, but also that authority is vested in the few (Heck, Bas, & Kinzler 2021). If children expected power differences to scale with size differences, they might also infer that stronger consensus would lead to faster decisions. Pun et al's (2016) studies were not designed to test whether power scaled with proportional differences (infants only saw groups of 3 and 2); but while Heck et al. (2021) did that find children and adults were more likely to attribute authority to proportionally smaller groups, their strength-in-numbers inferences did not scale with size. Taken together with Morgan et al., (2015), these findings suggest that reasoning about consensus strength and its effect on decision speed may involve capacities still developing between the ages of 6-9.

4.2 General method

In two pre-registered experiments, we tested our predictions by presenting children and adults with pairs of robotics teams deciding which of seven kinds of propeller would make a drone fly the best. In each trial, the two teams vary in the number of people, factions, or both. Participants are told that the teammates on each team will have to talk together to decide which propeller to use. They then rate how sure they are that one team or the other would take longer to decide on a seven-point scale (with the midpoint indicating no difference), and briefly explain their reasoning.

4.3 Experiment 1

4.3.1 Method

In Experiment 1, we asked children and adults to infer which of two teams would take longer to make a decision. Across three trials, we manipulated the number of people (*Size*), factions (*Diversity*), or both (*Contrast*). In the *Diversity* trial, two teams with the same number of people (10) were split into a different number of factions (2v7). In the *Size* trial, two teams with the same number of factions (2) differed in the number of people (10v20). In the *Contrast* trial, the team with more people (20v10) was split into fewer factions (3v7). We predict that both children and

adults will expect slower decisions from teams with more factions or more people, and that they will treat the number of factions as more important than the number of people (i.e., in *Contrast*). However, we expect these inferences to be specific to decisions: in a second task following the experiment (*Build*), we ask which of two teams (20v10) would take longer to *build* their drone, *after* a consensus decision had been agreed upon. In the *Build* trial, we predict that participants will expect a *smaller* team to take longer than a larger team: whereas the task of reaching consensus divides a team against itself, many hands may make light work once consensus is reached. The *Build* and *Contrast* trials also help rule out a simple "more is more" heuristic. If participants are simply mapping the "more time" response to the team with more people or more factions, they will expect no difference in decision speed when one team has more people and the other has more factions, and they will infer that that the *larger* team will take more time to build a drone.

Participants. We recruited 80 children in two age groups (40 age 6-7, M=6.95, SD=.50, and 40 age 8-9, M=8.98, SD=.58; 34 girls, no non-binary genders reported), as well as 41 adults through mTurk. One additional child fussed out before completing the experiment and was replaced.

Procedure. After practicing with the response scale, children were told that they would see two teams each making a remote control drone, but that the teammates disagreed about which of seven kinds of propeller (differentiated by the number of blades, from two to eight) would make the drone fly the best. The experimenter told the child that they would see "which kind of propeller *each* person on *each* team *thinks* is best", and that the teammates would need to talk together to decide which kind of propeller to use. The child's job was "to say *which* team you think will take *longer to decide* which kind of propeller to use". They were then shown three trials in one of four counterbalanced orders. In each trial, participants first saw a group of students, represented as silhouettes, divided into two teams (allowing for easy visual comparison of the total number of people on each team), and then were shown each teammate "standing next to" the propeller they thought was best. The experimenter then told the participant "*So now, all the people on the blue team have to talk together to decide which propeller to use. And all the people on the green team have to talk together to decide which propeller to use. But, which team will take longer to decide: the blue team, the green team, or will they take the same amount of time?*". Children were then asked whether they were "*just a little sure, pretty sure, or very sure?*"; adults responded directly on a sliding scale. Participants were then asked to explain why they thought that team would take longer to decide. Finally, at the end of experiment, participants completed one trial of a second task: they were told that the next two teams had *already decided* which kind of propeller to use, and all agreed — but now, they needed to build their drone. One team was shown to have 10 people while the other had 20 people; participants were told that each team would start building at the same time, and asked which team would take longer to finish *building* their drone.

4.3.2 Results

Results. Results are displayed in Figure 4.2. We conducted separate linear regressions on the child sample alone for each contrast *Type*, with responses centered on the midpoint of the 7-point scale and age in years centered on the midpoint of the children's age range (7.5 years), according

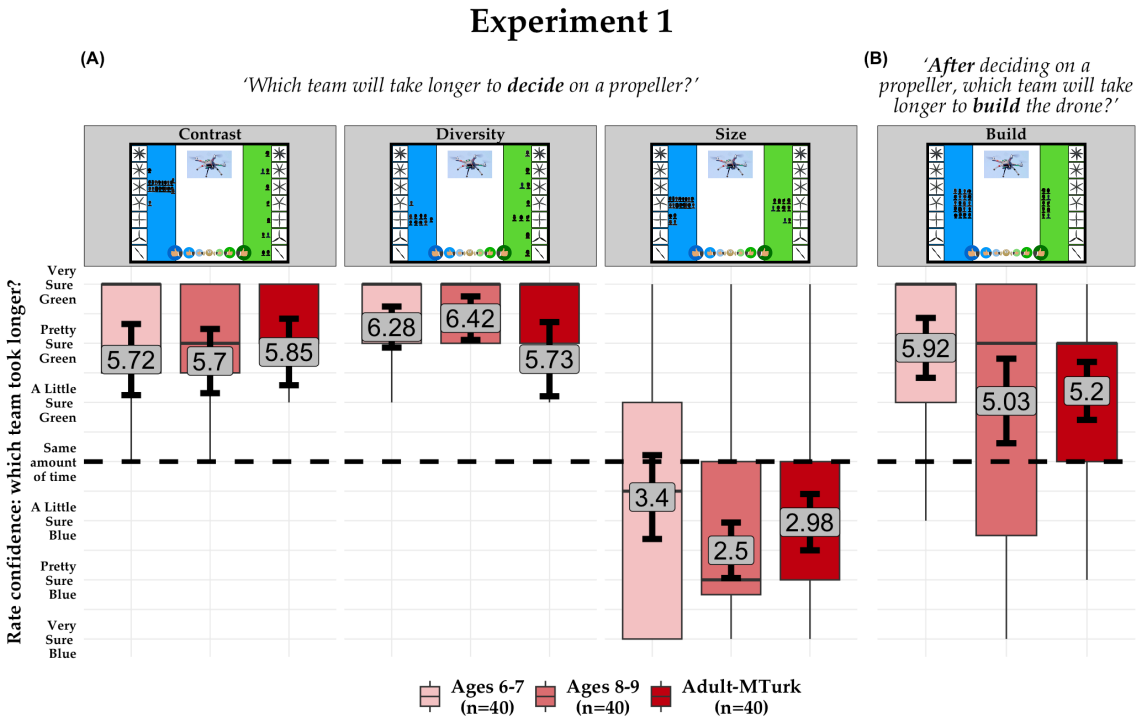


Figure 4.2a-b. Results from Exp 1. Shading indicates age group, grey labels display means, error bars are 95% CIs. Facets display critical slide from procedure. **(A) Decision time**: each participant rated each trial, in counter-balanced order. **(B) Build time**: each participant rated build time after completing all three decision trials.

to our pre-registered analysis plan. This makes the intercept equivalent to one-sample t.test versus the scale midpoint while allowing us to simultaneously account for potential age effects. There was no effect of counterbalance for any measure or age group, so we reduced the model to just Ct_Values ~ Ct_AgeYears for each contrast *Type*. As predicted, children expected slower decisions from a team with more factions than one with fewer (*Diversity*: $\beta_{Intercept}$ = 2.35, SE = .13, p<.001, *95CI*: 2.10–2.60); and when the team with more factions had fewer people, they expected the number of factions to matter more than the number of people (*Contrast*: $\beta_{Intercept}$ = 1.71, SE = .20, p<.0001, *95CI*: 1.31– 2.11). No age effects were observed in either trial. Moreover, the child sample as a whole expected slower decisions from teams with more people than fewer (*Size*: $\beta_{Intercept}$ = -1.05, SE = .21, p<.0001, *95CI*: -1.46– -0.64). But they also inferred that once the teammates had all agreed about their design decisions, a team with fewer people would take longer to *finish building* a drone than one with more (*Build*: $\beta_{Intercept}$ = 1.48, SE = .22, p<.0001, *95CI*: 1.04– 1.91). However, age effects were significant for both *Build* and *Size*, with older children more likely than younger children to infer that larger groups would take more time to decide, and less time to build (*Size*: $\beta_{Ct\_AgeYears}$ = -0.50, SE = .19, p<.01, *95CIs*: -0.87– -0.13; *Build*: $\beta_{Ct\_AgeYears}$ = -0.41, SE = .20, p=.039, *95CI*: -0.80– -0.02). Following our preregistered analysis plan, we also conducted one-sample t.tests comparing each age group (6-7s, 8-9s, and adults) to chance separately for each measure. Unlike adults and older children, the expectation of slower decisions from the larger team was not significant for the youngest children in the *Size* trial (M = -0.60, t(39)= -1.71, p<.095,

95CI -1.31—0.11). However, all other by-age-group t.tests were consistent with our primary analysis: all ages expected slower decisions from teams with more factions in the *Diversity* and *Contrast* trials, and slower builds from smaller teams in the *Build* trial.

What do these results tell us about participants' reasoning process? First, expecting the number of people on a team to have opposite effects on decision speed in the *Build* and *Size* trials shows that participants distinguished between the physical task of building and cognitive task of making a collective decision about how to build drone. Moreover, responses to the *Contrast* and *Build* trials demonstrate that participants were not simply mapping a "more time" response to the team with more people or more factions. Instead, participants' inferences about decision speed appeared to be driven by some notion of how the number of people and factions on a team affects the time needed to reach consensus. Roughly, this means reasoning about about some notion of disagreement, broadly construed; but there are a number of ways it could work that are less sophisticated than what we have in mind. For instance, one might simply assume *an* outcome (either majority rule, or whichever propeller seemed best to the participant themselves), an infer decision speed from the number of opponents remaining to be convinced. This is akin to the kind of reasoning predicted by our account, but because it's blind to differences in power that make some outcomes more likely than others, it will often generate counterintuitive predictions. For instance, one might expect convincing four people to always require the same amount of time, regardless of the number of factions and people in them (e.g., 16v4, 16v1v1v1v1, 2v4, 1v4, etc). In Experiment 2, we ask participants to infer which team would take longer given that *both* teams chose the optimal propeller. This allows us to control the numerical and proportional size of the winning and losing factions as well as the total number of people and factions, ruling out heuristics that assume a specific outcome or focus on a single numerical feature.

4.4 Experiment 2

4.4.1 Method

Experiment 2 probes participants' reasoning about how consensus strength affects decision speed. We predict that all ages will infer slower decisions from teams with more factions or people. But if the winning faction has less power relative to its opponents on a small team than the winning faction has on a large team, we predict that while adults will expect consensus strength to matter more than size, children will infer the opposite. For instance, adults might expect a minority rule outcome on a team of six to take longer than a majority rule outcome on team of twelve, children will infer the opposite. However, because Experiment 1 and the pilot data for Experiment 2 suggested that younger children's (ages 6-7) size inference may not differ from chance despite recognizing the impact of the number of factions, our preregistration treats older children as the primary developmental contrast for the trials in which size and factional power are contrasted. Younger may show the same pattern as older children; but if they do not differ from chance, further work would be needed to understand why.

Participants. Based on power analysis, we recruited 100 children in two age groups (50 age 6-7, M=6.88, SD=.67, and 50 age 8-9, M=8.98, SD=.67; 60 girls, no non-binary genders reported), as well as 50 adults through MTurk. Two children fussed out before completing the experiment and

were replaced; six adults were screened out and replaced before completing the experiment for failing an attention check.

Materials. We created four trials intended to contrast different dimensions of the distribution of opinions on each team: the size of each team, the number of options initially endorsed, and the proportion and number of teammates who had initially disputed the group's final decision. In two trials (*Maj_Min*, *SuperMaj_vs_Maj*), one team was twice the size of the other, but each team was split between two options, and choosing the correct propeller would require the team to convince 4 people to change their answer (*Maj_Min*: 8v4 or 2v4; *SuperMaj_vs_Maj*: 16v4 or 6v4). In the other two trials (*SuperMin_MinDiv*, *SuperMaj_PluralityDiv*), each team was the same size, but one team was split between all six options while the other team was split between only two options , with a either a plurality or majority initially endorsing or opposing the correct propeller (*SuperMin_MinDiv*: 4v16 or 4v6v3v2v2v1; *SuperMaj_PluralityDiv:* 16v4 or 6v4v3v2v2v1).

Procedure. The procedure was similar to Experiment 1, with the following changes. (1) First, during the introduction, participants were additionally told that "*the kind of propeller that's actually the best for the kind of drone these teams are both building is the one* 4-blades", after which the 4-blade propeller was highlighted in yellow and remained highlighted for the remainder of the experiment. (2) Second, after seeing during each trial what each teammate on each team thought was best, participants were prompted to remember which propeller was actually best. (3) Third, the experimenter told participants to pretend that both teams had ultimately chosen the correct propeller, saying: "*Now the teammates on each team have to talk together to decide which propeller to use. And each team might decide to use the 4-blade propeller, or they might not. And we don't know which propeller they'll choose after they talk. But, let's pretend we do know. Let's pretend that after they talk, both the blue team and the green team do decide to use the 4-blade propeller. So, which team do you think had to talk for longer, if both teams decided to use the 4-blade propeller: did the blue team take longer, did the green team take longer, or did they both take the same amount of time?*". (4) Finally, after rating how sure they were that one team or the other would take longer and explaining why, the experimenter told the participants "*Now we're done pretending for a minute. Remember, we don't actually know which propeller each team will decide to use — but, I want to know which propeller you think each team will use*", and for each team, asked the participant to predict whether the team would decide to use the 4-blade propeller after talking.

4.4.2 Results

Results. Results are displayed in Figure 4.3. Experiment 2 provides direct evidence against a number of heuristics simpler than the kind of reasoning about disagreement we have in mind. Across trials, children and adults made systematic inferences even when we controlled (1) the total number of people, (2) the total number of factions, (3) the number of "losers" (4) the proportion of "losers", and (5) the number and proportion of "winners".

On the *SuperMaj_PluralityDiv* and *SuperMin_MinDiv* trials, each team had 20 teammates; as predicted, children and adults expected slower decisions when they were divided into 7 factions than when they were divided into only 2 factions — not only when the team with more factions was contrasted with a team with a stronger winning faction (*SuperMaj_PluralityDiv*: 16-winners-

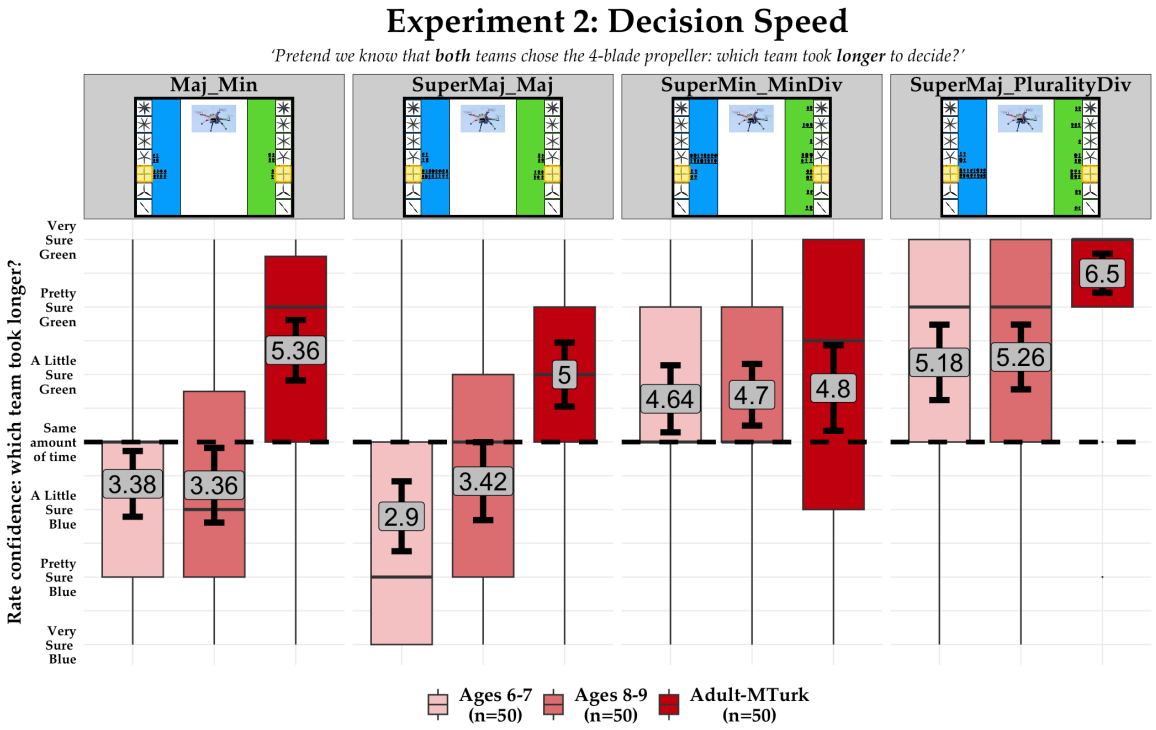


Figure 4.3. Results from decision speed task in Experiment 2. Participants were told that the propeller that was "actually the best" was the 4-blade propeller (highlighted in yellow), and told to pretend that both teams chose the best propeller after talking together. Facets display critical slide from procedure for each trial; each participant rated each trial, in counter-balanced order. Shading indicates age group, grey labels display means, error bars are 95% CIs.

vs-4-losers-in-1-faction and 6-winners-vs-14-losers-in-6-factions: $M_{younger}$= 5.18, $t(49) = 4.24$, $p<.001$; M_{older}=5.26, $t(49) = 5.28$, $p=.003$,, M_{adult} =6.50, $t(49) = 17.43$, $p<.001$), but also when contrasted with a team with the same number and proportion of both winners and losers (*SuperMin_MinDiv*: 4-winners-vs-16-losers-in-1-faction and 4-winners-vs-16-losers-in-6-factions: $M_{younger}$= 4.64, $t(49) = 2.59$, $p=.013$; M_{older}=4.70, $t(49) = 3.08$, $p=.003$,, M_{adult} =4.80, $t(49) = 2.54$, $p=.014$). One-way ANOVAs revealed that younger children were significantly less confident than adults on the *SuperMaj_PluralityDiv* trial; but older children's responses were not significantly different from either younger children's or adults' for either trial (*SuperMaj_PluralityDiv*: F(2, 147)=54.77, $p<.001$, $\eta_p^2 = .13$; *Younger—Adult*: $t(147)=-4.11$, $p < .001$; *Younger—Older*: all p's ns; *SuperMin_MinDiv*: F(2, 147)=0.65, $p=$ *ns; Older—Adult*: $t(147)=-0.27$, $p <$ ns; *Younger—Adult*: $t(147)=-0.43$, $p <$ ns).

On the *Maj_Min* and *SuperMaj_Maj* trials, each team was divided into 2 factions that left each team with the same number of "losers" to convince, but also made one team on each trial twice the size of the other (*Maj_Min*: 8v4-and-2v4; *SuperMaj_Maj*: 16v4-and-6v4). As predicted, adults inferred on both trials that the decision would have been slower when the winning faction was proportionally weaker, but children inferred that decisions would have been slower in the numerically larger team, even though the winning faction was proportionally stronger (*Maj_Min*: $M_{younger}$= 3.38, $t(49) = -2.56$, $p=.014$; M_{older}= 3.36, $t(49) = -2.33$, $p=.024$; M_{adult} = 5.36, $t(49) = 6.11$, $p<.001$; *Supermajority*: $M_{younger}$= 2.90, $t(49) = -4.27$, $p=.001$; M_{older}= 3.42, $t(49) = -2.02$, $p=.049$; M_{adult} =

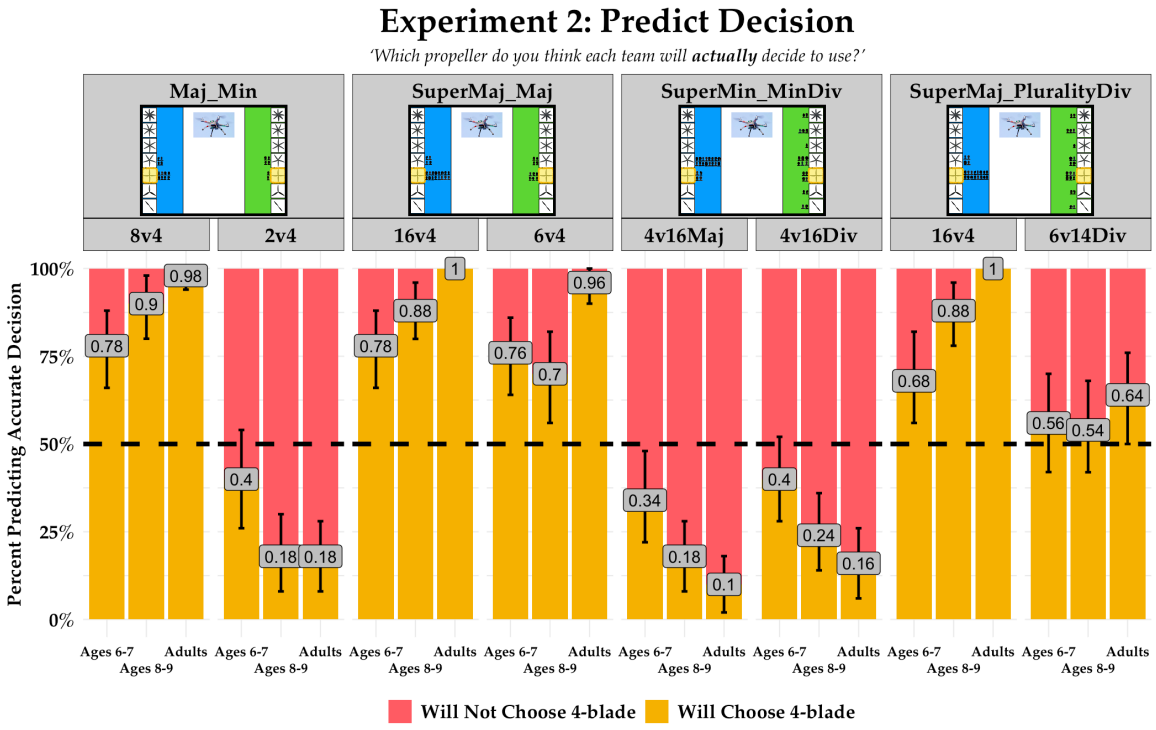


Figure 4.4. Choice predictions in Experiment 2. After inferring which team in each trial would have taken longer if both teams had ultimately chosen the 4-blade propeller described participants were told was "actually best" (highlighted in yellow), participants were asked which propeller they thought each team would *actually* choose, as a forced-choice between the 4-blade propeller (yellow) and any other propeller (red). Grey labels show means; error bars are 95% CIs.

5.00, $t(49) = 4.24$, $p<.001$). As predicted, these age differences were significant for older children (*Maj_Min*: *Older-Adult*: $t(147)=-5.71$, $p < .001$; *SuperMaj_Maj*: *Older-Adult*: $t(147)=-4.28$, $p < .001$). The pattern for younger children also differed from adults, but was indistinguishable from older children (*Maj_Min*: *Younger-Adult*: $t(147)=-5.65$, $p < .001$; *SuperMaj_Maj*: *Younger-Adult*: $t(147)=-5.69$, $p < .001$.

Since *SuperMaj_Maj* and *SuperMin_MinDiv* each contrasted two teams in which the winning faction was the initial majority, these results also speak against the possibility that inferences about decision speed are simply an artifact of assuming that only one of the teams (that without a majority) would need any time at all to make a decision. But when asked to predict each team's final decision, all ages expected majority rule — and to a lesser extent, plurality rule (i.e., in teams with many factions) — regardless of whether the propeller endorsed by the initial majority (or plurality) faction was the optimal decision or not (Figure 4.4). In other words, both children and adults *predicted* majority rule (and to a lesser extent, plurality rule), but their inferences about decision speed were not simply an artifact of assuming majority rule as a *fait accompli*.

4.5 General Discussion

Group consensus doesn't emerge solely from epistemic judgments. It's often negotiated, expedient, and costly to achieve. One cost is time. But collaborators who need to coordinate their decisions with each other no longer have unilateral control over the time they spend on a

decision or the decision itself. Instead, a decision's speed and accuracy both depend on social dynamics. And to manage speed-accuracy tradeoffs in groups, collaborators need to know how to pick their battles. Reasoning about endogenous constraints on group decision speed, such as the size and structure of a group, may allow collaborators to estimate the time needed to coordinate around one decision or another. Taken together, these two experiments suggest that some of the intuitions that may help us decide which battles are worth the time emerge in early childhood — but they may also undergo qualitative changes as a result of conceptual development.

Adults and children as young as six expected slower decisions from teams with more people or more factions. And these inferences were specifically about decisions: all ages expected that once consensus had been reached, drones would take longer to *build* in teams with *fewer* people to work on them. But while adults expected stronger initial consensus to speed up consensus-congruent decisions (and slow down consensus-incongruent decisions), children expected slower decisions from larger teams even when consensus was stronger than on the smaller team. We doubt children's size-over-strength inferences are due to a failure to realize that consensus was stronger on one team than the other. Even preschoolers can easily distinguish the vote ratios (2:1, 3:2, 4:1) we used in the two-faction trials (Halberda & Feigenson, 2008). And children did *predict* majority-rule, suggesting that they didn't have trouble recognizing the consensus preference — they simply didn't expect consensus strength to matter more than team size.

But why not? After all, group decisions aren't faster when consensus is stronger simply because of some arbitrary eccentricity. Agent-based models suggest that consensus strength is as much of an endogenous constraint on group decisions as the size of the group or the number of factions: lower decision thresholds (e.g., plurality or majority instead of supermajority or unanimity) and more impatient voters can speed up decisions, just as more people or more diverse preferences can slow them down (Albrecht, Anderson, & Vroman, 2010; Chan, Lizzeri, Suen, & Yariv, 2018). And these aren't just foibles of human decision-making. Other species encounter the same dynamics. When *temnothorax* ants urgently need to find a better nest, they lower their quorum threshold — enabling the "votes" of a smaller number of scouts to trigger a migration (Pratt & Sumpter, 2006). And when schooling fish choose a foraging patch, increasing the number of no-preference voters makes it harder for strong-preference minorities to overrule weak-preference majorities (Couzin et al., 2011; Ward et al., 2008). But other species' decisions are, presumably, less dependent on the kinds of metacognitive intuitions that make human judgment so flexible even among children; they don't adopt arbitrary rules to coordinate with collaborators or treat them as morally binding only for those who agreed to them (Grueneisen & Tomasello, 2019; Schmidt, Rakoczy, Mietzsch & Tomasello, 2016), and they don't use discussion to adjudicate ethical and rational dilemma (Domberg, Köymen, & Tomasello, 2019).

So why didn't children expect consensus strength to matter more than team size? We suspect a confluence of factors. As noted above, individual inefficiencies in communication add up quickly in large groups. And since children are much less skilled than adults at resolving conflict through meta-talk and reason-giving (Köymen & Tomasello, 2018), it wouldn't be unreasonable for them to expect slower decisions from larger groups in general; after all, when consensus is controlled,

so do adults. But while in adults, strength-dependent deference to consensus can short-circuit endless dissent and redundant commentary, this deference is only beginning to emerge around ages 6-7 (Morgan, 2015; Schmidt et al., 2016). Moreover, while children can explicitly justify the use of different decision-making procedures in different contexts (Helwig & Kim, 1999; Hok, Gerdin & Shaw, 2019), a more rigid sense of procedural justice (e.g., "everyone should have their say") could make majority rule as time-consuming as unanimous consensus.

Consider three examples of how the contrast between power and right could affect children's reasoning about the impact of consensus on social dynamics in group decision-making. First, when a group doesn't give a child opportunity to agree to their justification for distributing resources unequally, they are six to eight times more likely to object than if consulted first (Grocke, Rossano, & Tomasello, 2018). Consulting every member of a group in advance may take less time than handling their objections, but it would still make group decision-making more time-consuming in larger groups than smaller groups. Next, when groups agree to norms (e.g., about which puppets can play where), preschoolers treat dissent as nullifying the norm-establishment entirely — although they will occasionally protest if someone who agreed to a norm disregards it (Schmidt, et al., 2016). Our experiments examined a different context (artifact designs aren't arbitrary norms, and dissenters couldn't build their own drone), but if children are willing to allow single dissenter to veto a decision favored by nine other group members, they may not expect dissenters concede more quickly simply because they face a stronger consensus. And given that patience for dissent, children's inferences here may even accurately reflect their experience of group decision making: if stronger consensus doesn't put stronger pressure on dissenters to concede more quickly, group size may have a greater impact than consensus strength on decision speed. But lacking direct tests of children's collective decision times, whether or not children's inferences accurately reflect their experience — and whether developmental changes in reasoning about factional power could improve speed- accuracy tradeoffs in collective decision-making — are questions for future work. Finally, whereas adults believe that communities will count individuals as belonging to whichever subgroup the most powerful (larger, wealthier, and more prestigious) clique decrees — even without the individual's consent — children insist that even socially-perceived group identity requires consent from the individual themselves (Noyes, Gerdin, Rhodes, & Dunham, 2023). Our experiments examined opinion-based factions within a cooperative team instead of identity-based groups in a shared space; but if children are more respectful than adults of individuals' right to veto their socially-perceived identities, they may be also be more respectful of their right to make dissenting arguments — which would prevent consensus from speeding up decision making.

Most work on collective judgment has focused on its accuracy (Chittka, Skorupski, & Raine, 2009; Kameda, Toyokawa, & Tindale, 2022); and so has most work on children's strategies for learning from others (Harris, Koenig, Corriveau, & Jaswal, 2018). But good judgment isn't cheap: time spent improving accuracy is time lost for pursuing other goals, and the perfect may be the enemy of the good. Recent work has suggested that cost-reward reasoning may be fundamental to commonsense psychology (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). For instance, we not only expect others to rationally tradeoff expected costs against expected gains in pursuing

goals — we also infer what agents know and believe they can learn from the costs of action they're willing to pay (Aboody, Zhou, Jara-Ettinger, 2021; Aboody, Davis, Dunham, Jara-Ettinger, 2021; Aboody, Dension, Jara-Ettinger, 2021). But past work has typically quantified costs using physical dimensions (e.g., effort, distance traveled) or the risk of failure (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Time costs are more more ubiquitous than physical costs: every decision takes time, regardless of whether it involves movement or a probabilistic outcome. But they may also be more difficult to interpret (Richardson & Keil, 2022). Since time spent on a task is a matter of choice in ways physical costs can't be, time costs are more flexible; but whereas individual decision speeds are only constrained by the efficiency of biological processes and the complexity of the task itself, needing to coordinate collective decisions means that individuals no longer have (unilateral) control over speed-accuracy tradeoffs. Our experiments suggest that ceding *unilateral* control over time spent on collective decisions doesn't mean individuals cede *all* control: early-emerging intuitions about how group decision speeds are constrained by their size and structure may allow individual collaborators to weigh their beliefs about the value of different choices against easily-quantifiable constraints on the time it would take a group to converge on them. In some cases, expedience may be preferable to accuracy.

Of course, as the animal literature illustrates, stronger consensus can lead to faster decisions even without explicit representations of speed-accuracy tradeoffs, simply because the group is closer to its decision threshold from the outset (Conradt & List, 2009). And lacking the capacity to reason explicitly about those tradeoffs doesn't mean that coordinated collective decision-making can't be worth the costs (Miller, Garnier, Hartnett, & Couzin, 2013). But it's important to bear in mind that social dynamics in collective decisions are to some extent consequences of our beliefs about them: the more collaborators *expect* each other to concede more quickly to stronger consensus, the more pressure to do so dissenters may feel. These kinds of reflexive expectations can provide collaborators with a lever and a place to stand for more strategic inferences about each other's behavior. For instance, formal rules for quorum and decision threshold can make coordination easier, but they can also give swing voters disproportionate power, enable filibusters and agenda manipulation, and so on (Chan, Lizzeri, Suen, & Yariv, 2018; Levine & Plott, 1977; Pietraszewski, 2022; Pietraszewski & Shaw, 2015). Research into children's reasoning about factional power may help us understand the constraints on group dynamics. More broadly, the metacognitive capacities that make our inferences about complex social dynamics seem commonsensical may make us especially efficient at guiding collective action (Heyes, 2016).

Chapter 5

Conclusions

This chapter contains text from the following manuscripts:

Richardson, E., Hok, H., Shaw, A., & Keil, F. C. (*in prep*). Herding cats: Children's intuitive theories of persuasion predict slower collective decisions in larger and more diverse groups, but disregard factional power.

Richardson, E., Davis, I., & Keil, F. C. (*in prep*). Agenda setting and The Emperor's New Clothes: People infer that letting powerful agents make their opinion known early can trigger information cascades and pluralistic ignorance.

Richardson, E., & Keil, F. C. (2022). Anger, evidence, & trending opinions: We trust consensus when we believe it reflects genuine persuasion. *PsyArXiv*.

Richardson, E., Miro-Rivera, D., & Keil, F. C. (2022). Know your network: People infer cultural drift from network structure, and expect collaborating with more distant experts to improve innovation, but collaborating with network-neighbors to improve memory. *Proceedings of the Cognitive Science Society, 44*.

Cumulative culture highlights the extent to which collaborative and individual learning capacities mutually constrain each other. Collaboration allows cumulative culture to expand beyond our individual learning capacities; but as it expands, our individual capacities for skillful collaboration increasingly become the limiting factor in what we can learn either alone or with others. My central point here has been that skillful collaboration means thinking about how our collaborators' judgments are constrained by the interactions between them. This in itself is not a new idea. On the contrary: if thinking about those constraints didn't feel intuitive enough to be fairly frequently commented on, they probably wouldn't be much help in explaining our skill in collaboration to begin with. After all, the social dynamics collaborators face are both extraordinarily complex and very fast-moving; if they were also counter-intuitive, we wouldn't have much chance of handling them adaptively.

5.1 What happened in Chapters 2-4?

The experiments I've presented in this dissertation aimed to provide theoretically motivated descriptions of how some of the aforementioned commonsense intuitions emerge across development. I've intentionally avoided heuristics that require learners to have prior knowledge of differences between informants themselves (e.g., in expertise, past accuracy, or perceptual access). Instead, I focused on how children's and adults' inferences about the speed and reliability of a third-party judgment might be constrained by features more endogenous to those judgments: the nature of the problem itself, the number of informants facing it, the time needed to solve it given the means available, and the degree of consensus and dissent concerning its solution. I began, in the introduction, by talking about how these features hang together in a broad sense. What are the expected costs and benefits of deliberation for individuals and groups? How is the difficulty of a problem related to the means available for solving it and the time needed to do so? How does the perceived difficulty of a problem itself or the difficulty of resolving disagreement change the way we reason about the costs and benefits of private or collective deliberation? How

does the size and structure of a community affect the difficulty of solving a problem or resolving disagreement? Each chapter addressed overlapping subsets of these questions.

In Chapter 2, children and adults expected small discussion groups to be more help than "crowdsourcing" individual solutions for problems that could be solved through demonstrative reasoning, regardless of how difficult or easy they appeared to be — but a developmental shift in all three experiments suggested that while children expect larger crowds to be more helpful than smaller crowds, they may also either overestimate the benefits of group deliberation or underestimate its risks. In Chapter 3, children and adults inferred that the time needed to solve a problem when someone first encounters it depends on how complex it is, but the time needed to see or recall a solution is unrelated to its complexity. And in Chapter 4, children and adults infer that group decision speeds and accuracy both depend on the number of people and factions in a group; but only adults show clear evidence of reasoning about how the balance of factional power constrains a group's decisions.

5.2 What have we learned about the kinds of capacities that enable collaboration?

One of the take-home messages I'd like to endorse is that the kinds of inferences participants made in these experiments are critical to our capacity for learning from (and *with*) others — and thereby, to the remarkable pace of human cumulative culture. But the experiments I reported didn't provide measures of in-situ behavior; and while even children appeared to find the inferences commonsensical, it's possible that the intuitions I've examined in this dissertation don't make us better collaborators. After all, metacognition doesn't always guide behavior; when it does, it doesn't always make us better learners; and in any case, since the optimal learning strategies for individuals and their groups aren't always congruent or even compatible, one may be selected for more strongly than the other. Approaches that are more fine-grained and behaviorally-grounded than those adopted here could improve our understanding of how our intuitions about the risks and benefits of collaboration affect individual and collective learning. So before I summarize what I think I can say we *have* learned, consider some examples of nuances those conclusions may need to account for.

5.2.1 Limitations & closer looks

In Chapter 4, children expected slower decisions from groups with more factions or more individuals, but only adults expected stronger initial consensus to speed up decisions. These inferences are consistent with results of agent-based models showing that increasing a group's size and diversity slows down decisions, while lowering their decision threshold (e.g., plurality or majority instead of supermajority or unanimity) speeds up decisions. But those same models also show that when agents are able to represent their collective decision-thresholds, preferences for speedy or slow decisions take on a pivotal role: just a few impatient swing voters may give a stubborn minority enough leverage to overturn a moderate majority (Chan, Lizzeri, Suen, & Yariv, 2018). Overturning a majority may or may not make the decision more accurate, or maximize speed-accuracy tradeoffs; and even if it maximizes tradeoffs for collectives, it may not

do so for individuals. For instance, even though lowering their quorum threshold allows *temnothorax* ants (like humans) to make faster collective decisions between nests of similar quality, it does so in part because *individual* ants (like humans) still spend significantly longer evaluating nests that are more similar than nests that are more distinct (Pratt & Sumpter, 2006; Sumpter & Pratt, 2009; Sasaki, Stott, & Pratt, 2019). The point is that even if individuals' metacognitive inferences about collective dynamics *do* influence their collaborative behaviors, whether or not that influence makes them more better learners — or members of a collaboration — is a separate (and more complex) question.

Participants' inferences in Chapter 2 could also benefit from more nuanced analyses. The experiments in Chapter 2 were motivated by the observation that even though Condorcet-adjacent theories would predict more reliable consensus judgments from large crowds than small discussion groups, small groups often outperform both the average and majority response in large crowds, as well as their own best members (e.g., Mercier & Claidière, 2022, Navajas et al., 2018). And that performance gap is consistent with some of the theoretical advantages of information processing in groups (Hinsz, Tindale, Vollrath, 1997; Kameda, Toyokawa, Tindale, 2022). But the congruency between participants' inferences in Chapter 2 and specific empirical and theoretical advantages of groups and crowds is still too coarse-grained to make strong claims about how adaptive our commonsense intuitions really are. For instance, while the preferred response to the "non-reasoning" questions in each experiment (i.e., challenging perceptual discriminations or population preferences) shifted across development from a five-person discussion group to fifty-person crowd, the more reliable option for those questions may not always be the large crowd. Why not?

Recall that Condorcet's assumption of statistical independence between judges is routinely violated in the biological world, even if judges don't influence each other directly: judges often share cognitive and perceptual biases as well as a learning environment, which means they'll often rely on the same sources of evidence and interpret that evidence in similar ways. If a perceptual task falls afoul of our biologically inherited perceptual biases, relying on consensus in a larger crowd will simply amplify their mistakes. The same thing can occur when (as is typical in shared learning environments) judges each use a mix of jointly- and independently-observed cues which are themselves more or less correlated with each other and vary in reliability. Suppose that, unbeknownst to you, the burgers at Bobcat Bite are likely to be great on days when Greta's in the kitchen (M-W-F) and "merely" good on days when Gordon is (T-Th-S). If you ask your friends about about Bobcat's burgers, how well their consensus judgment predicts whether or not you get a great burger clearly depends less on whether they all went on the same day than on *which* day(s) people are most likely to go. But it also turns out that when your friends *are* more likely to have gone on the same day, you'll often be better off asking *fewer* of them about their experience: larger crowds will more accurately reflect the quality of the burgers coming out of the kitchen that day, but since Greta and Gordon work equally often, larger crowds will also underestimate the odds of getting a great burger and overestimate the odds of getting a merely good burger (Kao & Couzin, 2014). After all, a crowdsourcer is simply inheriting the crowd's biases, and a larger crowd just amplifies those biases. But if your friends have the right kinds of

skills, allowing them pool their evidence in groups offers a variety of ways to make their judgment more reliable, even if they can't engage in demonstrative reasoning. In some contexts, you don't even need to assume those skills include language (Kao et al., 2014). If learners' confidence is calibrated to their accuracy, pooling evidence can mean simply sharing metacognitive signals (Bahrami et al., 2010); it will still be sufficient to make two heads better than one. Alternatively, suppose that all learners can decide for themselves is whether to base their own vote on a jointly-observed or an independently-observed cue, and they can only see the payoffs of following the cue endorsed by a majority (but not the payoffs of the unendorsed cue); in this case, standard reinforcement learning algorithms allow learners to identify the more reliable cue more quickly and accurately than if they learn the associations individually and their votes are artificially aggregated at each round (Kao et al., 2014).

In short, strong generalizations about the value of consensus-based learning strategies not only need to consider consensus' reliability in light of the ground-truth dependencies for a specific question (e.g., Prelec, Seung, & McCoy, 2017), they also need to consider whether consensus outperforms *alternative* strategies — and if so, why. For instance, eliciting independent judgments from large crowds can be time-consuming; but, as Chapter 4 shows, even children know that group deliberation can time-consuming as well. And speed-accuracy tradeoffs are just one dimension on which to compare group- and crowd-based learning strategies. One might also contrast their costs and benefits in social cohesion, demands on individual effort (Miller, Garnier, Hartnett, Couzin, 2013), or their robustness across learning environments that change over time (Boyd & Richerson, 1995).

Finally, the phenomena studied in Chapter 3 exemplify the same kinds of concerns about speed and accuracy discussed above, but at the level of individuals instead of groups. Given the complexity of the puzzles used, one could argue that some responses in Experiment 2 (i.e., 20s trials for complex maps) and Experiment 3 (i.e., 3s trials for complex maps) suggest that both children and adults are wildly underestimating the time needed to accurately solve a complex problem. But if we expect others' intuitions about speed-accuracy tradeoffs to be similar than our own, then the speed of their response tells us something about the degree of accuracy they think is *worth* trying to achieve. The social learning literature has emphasized that blind trust in others' judgment is unlikely to be adaptive because of the potential costs in *accuracy*; but skepticism has costs too — in *time* (among other things). If Alice can't trust Bill's assessment of the speed-accuracy tradeoffs he faced in a given problem, double-checking his work may cost her more time than collaborating with him saved her. These kinds of intuitions may be critical to the phenomena discussed in Chapter 4, especially as groups grow larger: a conversation full of pedants won't simply drag on far longer than it needs to — it will collapse under its own weight as people lose the thread of conversation amid the quibbles and clarifications.

5.2.2 Developmental changes: learning to reason about reasons?

Each chapter in this dissertation suggested significant changes across development. The most revealing patterns may be the increasing preference for crowdsourcing over small group discussion in Chapter 2 (particularly in Experiment 3) and the contrast between children's belief

that team size would affect decision speed more than initial consensus strength and adults' opposite belief in Chapter 4.

One explanation that may account for both patterns is changes in children's understanding of the *non*-epistemic reasons people form (and change) beliefs, and the errors and biases those reasons can introduce. In each experiment in Chapter 2, the developmental shift towards crowdsourcing only affected *non*-reasoning questions; the tendency to refer reasoning questions to small group discussion was essentially identical across age groups and experiments. This shift may be particularly revealing in Experiment 3, where the youngest children preferred small group discussion for both reasoning and non-reasoning questions. Why? Increased crowdsourcing for the non-reasoning questions in Experiments 1 and 2 could be due to children's increasing awareness that crowdsourcing *is* the method for determining population preferences; this would explain both the developmental shift in Experiments 1 and 2 and participants' greater confidence in a 50-person crowd (Experiment 2) than a 5-person "crowd" (Experiment 1). But in Experiment 3, the preference for discussing difficult perceptual judgments couldn't be explained by their difficulty (since the preference was *stronger* for the reasoning questions, which were easier according to a pre-test); and so children may have either overestimated the discussion group's ability to discern accuracy without demonstration, or underestimated the potential for discussion to distort perceptual judgments. Children's responses in Experiment 2 of Chapter 4 may reveal a similar kind of reasoning: they may have inferred that consensus wouldn't speed up decision-making because they believed that even if majority rule ultimately prevails, it's not because people who found themselves in the minority immediately conceded — people have to be convinced, or at least have a chance to disagree. The same story could be told about Experiment 2 in Chapter 3: the youngest children may have been less confident than adults that too-fast responses to difficult questions were wrong because it would require them to explain why an agent was claiming knowledge he simply hadn't had time to acquire. In short, one possibility for a general account of children's responses is that, roughly speaking, they expect people's beliefs to be more influenced by good reasons than bad reasons, regardless of whether these influences are social or asocial.

Changes in our confidence in other people's discernment (i.e., being more compelled by good reasons than bad reasons) would be consistent with a commonly observed developmental bias: we're more likely to attribute true beliefs to others than false beliefs. This may be an adaptive bias for learners who are as deeply dependent on others' knowledge as children are, particularly if they can count on their informants' discernment of each other's knowledge. But it would also raise questions about how we learn to update our beliefs when an informant's reliability is put in doubt. For instance, 4-year-olds trust a 3-to-1 consensus when they hear Bob and Carol repeat Alice's whispered testimony about the contents of a box after she and David each look inside; but they reject the consensus when Alice had announced that she was going to "pretend" to know instead of looking before whispering to her friends (Kim & Spelke, 2020). But reasoning about perceptual access is a familiar, firsthand competence. The rational update for a layperson might be less clear if Alice announced that her statistical analysis of racial disparities in police shootings controlled for differences in the number of encounters — phenomena such as Simpson's paradox

and collider bias require culturally transmitted conceptual systems for evaluating evidence (Ross, Winterhalder, & McElreath 2018). The division of cognitive labor means that learners have to outsource evaluating their informants' reasons on the merits to still other informants. But exposure to the explanatory preferences of experts in different fields may help us learn to discern more abstract features of good and bad reasons; it may also help us discern whether Alice's mistakes are a result of biases, bad reasoning, or general lack of expertise.

5.2.3 Beyond demonstrability

The three sets of studies in this dissertation all focused on how people reason about agents facing problems with more-or-less demonstrably correct answers. Though humans' capacity for solving these kinds of problems is thought to be critical to cumulative culture, focusing on that capacity also omits the greater part of human life: as social animals, we're usually much more occupied by coordination, convention, and arbitrary norms than we are by solving high-demonstrability problems. How might participants' inferences in Chapters 2-4 change if agents were facing decisions like where to meet for lunch, which side of the road to drive on, or which job candidates will bring the most prestige to an organization?

To the extent that solutions to these kinds of problems aim for consensus rather than "truth", and consensus can emerge *without* deliberation (e.g., through stigmergy, self-organization, silent conformity), one possibility is that people will see the costs of deliberation as more salient than the benefits. However, the lack of a truth criterion may also reduce consensus about *which* costs and *whose* costs are most relevant, and since deliberation becomes impossible in larger groups, choosing to deliberate may also entail decisions about whose costs are represented in the deliberation to begin with. For instance, a silent vote or even an executive decision can tell people where to meet for lunch with less time and effort than deliberation; but they're not as good at accounting for costs like conflicts between one person's food allergies and another's dietary preferences. Similarly, the costs of driving on the wrong side of the road are too high to allow consensus to emerge spontaneously, but since the optimal solution may need to satisfy too many constraints to entrust to a general vote, decisions may need to be delegated to a deliberative committee or an expert planner— in other words, to a (more-or-less representative) government. And since groups typically default to majority rule when demonstrability is low (Laughlin, 2011), participants might also assume that agents' deliberation would be more strongly shaped by deference to majority preferences.

The broader point here is that the greater diversity of potentially relevant costs in problems without a truth criterion may make people's reasoning about the experimental contrasts much more dependent on their scenario-specific inferences than high demonstrability problems are. For instance, participant in Chapter 2 might have inferred that five people choosing a lunch spot for themselves should discuss instead of voting (in Experiment 1). But if she believes the decision is being made for a larger group (i.e., because Experiments 2 and 3 contrast a 5-person discussion with a 50-person crowd), or if she's just seeking recommendations for herself, she may choose the 50-person vote because of concerns about representation (either in the sense of giving people a voice in decisions that bind them, or in the sense of avoiding sampling bias in crowdsourced

reviews). The experimental procedure in Chapter 4 makes concerns about representation and sampling bias less relevant, but if people expect majority preferences to more strongly influence decisions about lunch spots than decisions about drones, even children might treat factional power as more important than team size. And in Chapter 3, an agent's response speeds to questions about arbitrary conventions may still reveal something about the its complexity or the agent's familiarity with it, but questions about the agent's "accuracy" lose meaning.

5.3 Cognitive systems at the group level

Where does this leave us? It's worth considering the implications that early-developing intuitions about collaborative learning may have for a fundamental question in the cognitive sciences: what counts as a cognitive system? On the one hand, cognitive scientists are ostensibly concerned with describing and explaining such systems. On the other hand, asking a cognitive scientist to explain what makes a system *cognitive* may be embarrassing; there is no consensus account, nor a single criterion for evaluating them (Adams & Aizawa, 2001). What does this debate have to do with our intuitions about collaborative learning?

Briefly put: those intuitions may help us understand the nature of cognitive systems at the group level. Here's the bones of the argument (I'll flesh it out in a moment): if certain hallmarks of human cognitive functioning perform best in groups, and show characteristic error patterns when tested in isolated agents, these error patterns may mark the joints of distributed cognitive systems. The division of cognitive labor that we observe in the storage of acquired knowledge and in the acquisition of new knowledge arguably provides these marks. And an early-developing preference for group over individual processing — one accompanied by a suite of intuitive theories producing systematic inferences about how to minimize errors in the storage and acquisition of different domains of knowledge — may imply that human cognition is built to take advantage of distributed processing when appropriate.

5.3.1 Carving the (collective) mind at its joints

A fruitful approach to psychological research has been to "carve the mind at its joints". But over the last twenty years, the extent of the mind itself has been subject to intense debate, fueled by research suggesting that distribution of cognitive labor across multiple agents may be as essential to memory (Wegner, 1987; Theiner, 2013; Mahr & Csibra, 2018) and learning (Goldstone & Theiner, 2017; Keil, 2006; Harris, 2002) as it is to technological development (Kitcher, 1990, 1993; Mesoudi, 2017). One view has proposed that the mind should be considered to include any process that it necessarily relies on to function, such that, were it inside the head, we would consider it to be part of the mind — for example, an amnesiac and the notebook he uses to replace his memory (Clark & Chalmers, 1998). Critics retort that this criterion leads to "cognitive bloat" and a loss of explanatory power. After all, physiological adaptions have also transformed humans into obligatory cooks: even when using blenders to reduce the load on teeth and the digestive system, raw foodists find it difficult to eat enough volume or variety to maintain health. Yet, our dependence on frying pans and blenders doesn't lead us to speak of "extended digestion" (Sterelny, 2010). Rather, digestion is explained primarily by the internal organs, just as

the amnesiac's memory is primarily explained by the amnesiac's cognitive processes rather than the properties of the notebook. Critics who take this view would predict that even if humans frequently rely on artifacts to scaffold cognition, perception, and memory, it's unlikely that cognitive science will discover interesting regularities governing human-artifact interactions (Adams & Aizawa, 2001; 2008).

Like others (e.g., Theiner, Allen, & Goldstone, 2010; Sterelny, 2010; Heersmink, 2015), I take this to be a helpful critique of many human-artifact interactions in many respects. But if, as seems to be the case, an individual human mind is sufficiently modular to be "carved at the joints" in a way that allows both the modules and the mind as a whole to still be considered paradigmatic cognitive systems, then the objection to human-artifact systems (namely, that their "joint" capacities are explained entirely by the capacities of the human mind) doesn't seem to apply. This means that at least some kinds of cognitive systems can be linked in ways that produce a *new* macro-system, with capacities that neither component has. The individual human mind, consisting of coupled modules, is one such macro-system; but human groups, particularly in collaborative problem-solving, may be another. However, on this view, the explanatory focus needs to change: since cognitive systems often interact with each other, saying that all interacting cognitive systems are macro-systems doesn't tell us anything new. Instead, we need to understand what kinds of links allow macro-systems to gain capacities the component-systems didn't have, and what kinds of capacities the component-systems need in order to manage their interdependence.

One approach to these questions has been to focus on the kinds of mechanisms and phenomena that are relevant to a broad range of "linked" systems: network size and structure, connection bandwidth, and how these features affect the efficiency and reliability of diffusion and inhibition processes (Goldstone & Theiner, 2017). These are common manipulations in research on cumulative culture. For instance, while a community can simply be too small to maintain a broad knowledge base (Kempe & Mesoudi, 2014; Derex et al., 2013), subdividing sufficiently large social networks into smaller clusters can increase its knowledge base by changing explore-exploit decisions (Derex & Boyd, 2016). Though conformist tendencies may still influence individual agents' exploration patterns *within* clusters, *between*-cluster influences are reduced, allowing clusters to drift apart. Restoring the lines of communication between clusters then allows them to combine what they've learned, producing innovation. In "rugged" fitness landscapes, which contain multiple good-but-not-optimal solutions, fragmenting networks can thus increase learning by encouraging individual learners to explore more diverse options (Mason, Jones, & Goldstone, 2008). Simulation studies and in-lab experiments suggest that these manipulations can dramatically increase the speed of cultural accumulation. Moreover, in at least some domains technological improvements can accumulate over time even if individual agents have no understanding of the causal mechanisms underlying the technology they're developing (Derex, Bonnefon, Boyd, & Mesoudi, 2019).

5.3.2 Local errors, local illusions, and protocols for collaborating in social networks

But, at least in the case of human groups, another approach may be to rely on some of the same kinds of techniques used to carve the individual mind at the joints — with some modifications: if individuals are being studied in isolation from groups, then to some extent, the joints have already been carved — the point is to understand how they fit back together. Consider three examples of how this kind of approach might lead to reinterpretations of existing work, and provide a framework for further study.

5.3.2a "Lesioning" groups.

Lesions to individual brains produce systematic errors or a loss of capacities that rely on interdependence between the systems, but still allow the component systems to perform cognitive functions for which that interdependence was not causally relevant (something that, at least prior to the last decade, most artifacts no longer did after being separated from their users). If a cognitive system is distributed across multiple agents, one might expect individuals to make systematic errors in isolation that they don't make when collaborating with a group — for tasks that involve that system. Group advantages for certain kinds of reasoning and memory tasks appear to show precisely those kinds of errors. Evidence I reviewed in Chapters 1 and 2 suggests that small group discussion allows learners to solve reasoning problems that prove to be impossible to solve individually, even for the group's best members (e.g., Laughlin, Bonner, & Altermatt 1998; Moshman & Geil, 1998). But forcing individuals to work alone — i.e., carving them out of their social networks — doesn't reduce performance across the board. Individual reasoners are very good at evaluating arguments they disagree with; it's their ability to produce arguments for their own positions that appears to be lazy and biased. In other words, their errors are systematically biased, and biased in precisely the way you would expect if individuals were preemptively outsourcing responsibility for critiques to group members who disagreed (Mercier, 2016). Accounts differ on whether these biases have genetic (Mercier & Sperber, 2011; 2019) or cultural roots (Heyes, 2019; Dutilh-Novaes, 2020), but both emphasize the distributed nature of reasoning.

Research on transactive memory systems has prompted similar arguments. Intimate couples and members of naturally occurring groups are often aware of each partner's domain of expertise, and outsourcing responsibility for that domain to a partner allows them to outperform pairs or groups of strangers and their own individual performance (Wegner, 1987). Moreover, assigning a memory strategy that conflicts with their natural division of labor harms intimate couples' performance, but helps strangers' performance. That is, simply having agents interact doesn't make them into a single cognitive system distributed across two minds. But some transient groups do form transactive memory systems (Liang et al., 1995), and the more that the task actually *necessitates* cooperation between group members in order to succeed, the stronger this system becomes (Harris, Barnier, & Sutton, 2011; Brandon & Hollingshead, 2004; Theiner, 2013). As with reasoning, recent accounts have suggested that some aspects of memory (e.g.,

episodic memory in particular) are naturally distributed among social partners (Mahr & Csibra, 2018; Theiner, 2013).

5.3.2b Illusions of knowledge.

Visual illusions are phenomenological experiences in which gaps in the information available to a system are filled in by its best guess about what *should* be there**.** But our sense of how much of our knowledge is "in the head" is often illusory as well. For example, individual speakers can successfully use words like "beech" and "elm" to communicate with each other about their intended referent, even though they're unable to distinguish the two trees for themselves. On a strong externalist account, the meanings of these words are simply not stored in our heads at all (Putnam, 1975); on an internalist account, the meanings *are* in *a* head, just not the *speaker's* head — the speaker's words refer to concepts that are stored in the head of an expert who *can* distinguish between them (Jackson, 1998). But on either account, the cognitive system that supports the "meaning of words" is distributed. However, it's often the case that considerable torque has to be applied to our phenomenological experience before we realize how much of what we know "ain't in the head". For instance, in one study (Kominsky & Kei, 2014), laypeople overestimated the number of differences they can name for words whose referents are easily distinguishable to experts (ferret-weasel, dinner-supper, or cucumber-zucchini), but not for synonyms or word pairs with well-known differences (e.g., dog-wolf, or baby-infant). Similarly, laypeople are more likely to claim that they personally understand a novel phenomenon when told that scientists have published a complete explanation of it than when told that scientists did not yet understand it themselves, or when told that the scientists' explanation was a classified state secret (Sloman & Rabb, 2016).

But importantly, illusions of knowledge aren't just egocentric biases; they can be amplified by collaboration and attenuated by re-establishing social contexts (e.g., observing others, or explaining without access to collaborators). For instance, in one study (Richardson & Keil, 2021), I showed that if children are "scaffolded" with subtle hints while trying to figure out how to use a complex mechanical artifact that 0% of unscaffolded children learned to use alone, 93% of scaffolded succeed; but even among 9-10 year olds, only 35% recognized that they *needed* the help. However, after observing a third-party being scaffolded, recognition that scaffolding was necessary doubled. Similarly, children and adults often overestimate how well they understand the details of complex artifacts and other complex causal systems; but asking them to provide step-by-step mechanistic explanations — the kind you would have to give to demonstrate your understanding to other people — reduces their confidence considerably (Rozenblit & Keil, 2002; Mills & Keil, 2004).

5.3.2c Commonsense intuitions about collaborative learning.

Systematic errors and illusions of knowledge could mark the lesioned joints of a distributed cognitive system. But unlike modular systems in a single brain, our groups aren't hardwired into a single biological interface: other people have their own beliefs, goals, and social networks to maintain. Recruiting assistance from a group won't be worth Alice's time and effort if it turns out that Bob and Carol flatly refuse to cooperate with each other, or if David absconds with a subset

of the knowledge and skills the group needed to accomplish a task, or if no one is willing to disagree with Alice herself about anything. The point is, collaborators need to be able to manage their interdependence with other parts of a distributed system in order for those systems to do any work. But the kinds of problems collaboration excels at solving are sufficiently varied and complex that the kinds of pre-compiled evolutionary programs that drive cooperative behavior in other eusocial species are unlikely to work for humans. The human capacity for collaboration may need to be guided by introspectively accessible metacognitive intuitions. By analogy: if computers were considered cognitive systems, we might ask what specialized programming allows them to form networks, and to what extent they require certain types of networks to accomplish certain kinds of tasks. Early-developing commonsense reasoning about how our collaborations are constrained by speed-accuracy tradeoffs and the mutual influences between our collaborators and their communities may be one kind of "network protocol" for managing our distributed capacities. To be clear, the kind of intuitive reasoning I've focused on in this dissertation may improve our capacity for collaborative learning even if it turns out that there are no group-level cognitive systems. And as noted above, much more empirical data and theoretical development would be needed to show what explanatory power could be gained by treating individual minds as components of group minds.

5.4 Conclusions

Humans are a spectacularly successful species. Though cumulative culture has advanced technology beyond what any individual could learn on their own, children and adults are adept at identifying reliable sources to learn from using a variety of cues to expertise. However, humans also learn *with* others, and by adulthood, we form our social networks into ad hoc groups, long term collaborations, and institutional structures — each governed by a variety of formal and informal decision rules. Research emerging in the past twenty years has suggested that the size and structure of these groups play a significant role in cumulative culture; but their impact on children's learning and the individual psychological capacities that allow us to learn from them are still unclear.

Technological developments in the last 10 years appear to have radically expanded our social networks, and underline the need for cognitive scientists to study information processing in groups, including its developmental origins. Children's intuitive theories of group cognition, and their ability to form groups of different structures to solve different kinds of problems, may provide a framework for doing so. This dissertation has presented evidence for early-emerging commonsense intuitions about the costs and benefits of using groups and crowds to process information and solve problems that exceed their individual capacities. I've also given a sketch of how these intuitions could shed light on the architecture of group-level cognitive processes.

Appendix A

Supplemental Materials

A.1 Supplement to Chapter 2

A.1.1 Experiments 1 & 2: Comprehension Questions

After the test questions in Experiments 1 and 2, we asked two comprehension questions ("Comp_TT" and "Comp_AA") to test more explicitly whether participants were considering the effects of information sharing in a setting familiar to children. In these questions, Jack's teacher was giving a test to Jack's 5 informants, and participants were asked whether the 5 people should answer by Talking Together or by Answering Alone. In Comp_TT, the teacher wanted "the 5 people to get as many answers right as possible"; in Comp_AA, the teacher wanted to "find out which of the 5 people did their homework and which ones didn't". If children understand how discussion changes the informativeness of individual responses, they should recognize that Answering Alone is more informative to the teacher in Comp_AA. If they understand the benefits of discussion (or at least, information sharing), they should prefer Talking Together for Comp_TT.

In Experiment 1, children's responses to the comprehension questions suggest that even the youngest were able to choose a method of responding consistent with what the teacher wanted to learn about the students (Comp_AA: M_{Young}= 65%, p=.04, M_{Old}= 87.5%, p<.0001, M_{Adult}= 92.5%, p<.0001, Comp_TT: M_{Young}= 70%, p=.008, M_{Old}= 85%, p<.0001, M_{Adult}= 87.5%, p<.0001).

As in Experiment 1, responses to the comprehension questions at the end of Experiment 2 suggested even the youngest children recognized that talking together would make it impossible for the teacher to know who had done their homework (Comp_AA: M_{Young}= 67.5%, p=.019, M_{Old}= 92.5%, p<.0001, M_{Adult}= 90%, p<.0001). However, while older children and adults recognized that the students would do better on the test if they could discuss their answers, younger children were at chance (Comp_TT: M_{Young}= 52.5%, p=.4373, M_{Old}= 90%, p<.0001, M_{Adult}= 90%, p<.0001). Children in Experiment 2 may have been less confident in the value of discussion than their responses to the the main task questions in Experiments 1 and 2 would suggest; however, informal questioning of participants after the experiment suggested that younger children in Exp 2 may have simply rejected talking together on a test as cheating, even though the question specified that the teacher themselves could choose to allow students to talk together.

A.1.2 Supplementary Methods for Experiment 3

Norming Experiment. In order to confirm the difficulty level of the Hard Percept and Easy Reasoning questions in Experiment 3, we first ran a norming experiment on MTurk with a separate group of 42 adult participants. Three participants were screened out for failing to answer basic comprehension questions about their job in the HIT.

We created 8 questions (4 Percept and 4 Reasoning) that we expected participants to rate as "easy" to answer and another 8 questions (4 Percept and 4 Reasoning) that we expected participants to rate as "hard" to answer. Each participant saw 8 questions: either the 4 Easy Reasoning and 4 Easy Percept questions, or the 4 Hard Reasoning and 4 Hard Percept questions. We expected the Hard Percept questions to be rated as more difficult to answer correctly than the Easy Reasoning questions. Each participant was asked "How difficult would it be to answer the question?", and rated the difficulty on a 7 point scale, from *Extremely easy* to *Extremely difficult*.

The *Percept* questions:

Photorealism: decide which of two pictures of a face is a photo and which is a photorealistic drawing made by a talented artist. These materials adapted from Looser & Wheatley, 2010, which morphed faces using photographs and dolls as the anchors. We used Morph 3. The Easy version used Morph3_052Human and Morph3_067Human. The Hard version used Morph3_063Human and Morph3_065Human.

Intuitive Psychophysics (Superballs): decide how many marbles an opaque box contains by listening to it being shaken. This task was adapted from Siegel, Magid, Tenenbaum & Schulz, 2014. Two recordings were created. The Easy version asked whether the box contained 2 or 10 marbles (the recorded version contained 2). The Hard version asked whether the box contained 30 or 40 marbles (the recorded version contained 40).

Brightness (Stars): decide which of the stars in a starry night sky looked the brightest. A picture of a starry night sky over a desert was used to represent the night sky, and the protagonist was said to have taken the picture so that he could "circle the brightest ones". In the Easy version, he wanted to circle the 3 brightest stars. The Hard version he wanted to circle the 25 brightest stars.

Rotation Speed: identify which of twelve colored diamonds is rotating the fastest. Each diamond had an A, a K, or a W in it to make the rotation clearer, but in the Hard version, the diamonds all had approximately the same RPM, while in the Easy version, the RPM was overall slower, and one was a clear outlier. The matrixes below show the number of rotations of each item in the Hard and Easy 4x3 arrays during the 10s display. In the Hard array, the fastest made 27 rotations in 10s, but 3 others made 26 and 2 made 25 rotations. In the Easy array, the fastest made 19 rotations in 10s, and the next closest made 12.

Hard (# Rotations/10s)	Easy (# Rotations/10s)
24 22 21 **27** 26 26 26 25 25 22 25 22	8 6 10 7 6 3 7 **19** 8 9 12 11

The *Reasoning* questions: The reasoning questions were adapted from Experiments 1 and 2.

Sudoku: Experiments 1 and 2 used a 4x4 sudoku problem rated as "easy" in a compilation, replacing the numbers with fruit to make it kid-friendly. The Easy version in Experiment 3 completed two additional moves. The Hard version used a 9x9 rated as "hard" in a compilation.

Vehicle Routing Problem: Experiments 1 and 2 used a custom made pathfinding puzzle which required a MarioKart find the shortest road through all the treasures on a map without taking "two in a row that are the same color, or two in a row that are the same shape". The Hard version used in these experiments had 11 treasures of different shapes and colors scattered randomly around the map. The Easy version created for Experiment 3 reduced the number of treasures to 4, of only 3 shapes and colors.

Bottle-Jar Extraction Task: Experiments 1 and 2 presented an "impossible object" puzzle, requiring the solver to remove a stick from a bottle without breaking the bottle or the stick. The stick was held fast inside the bottle by a nut-and-bolt. This was used as the Hard version. The Easy version substituted an analog of the "floating peanut" task (e.g., (Hanus, Mendes, Tennie, & Call, 2011), requiring the solver to remove a rubber ducky from large open-neck jar half-full of water, without touching the ducky or the jar, by pouring in the water from another jar.

Nim: In the game of Nim, each side takes turns picking up pencils. Each turn, you have to pick up either one, two, or three pencils. The winner is the person who picks up the last pencil. In Experiments 1 and 2, the we showed a game with only 5 pencils left. As adults and some older children found this 5-item version easy to solve, we created a Hard version by leave 22 pencils, and emphasizing that a wrong move would let a "super-smart computer" opponent win.

Norming Experiment: **RESULTS**. We fit a mixed effects model to perceived difficulty ratings, with random slopes and intercepts for each participant and question to account for repeated measures. The model confirmed that participants expected the Hard questions to be more difficult to answer than the Easy questions, ($\beta = 1.95$, SE = .5523, $p = .0055$). With the exception of the Easy version of the Percept_Stars question, which was rated as significantly more difficult than other Easy questions ($\beta = 2.45$, SE = .0.4988, $p < .0001$), the questions within each difficulty level did not differ amongst themselves in perceived difficulty. Experiment 3 contrasted the Easy versions of the Reasoning questions with the Hard versions of the Percept questions; if participants preference for group reasoning in Experiments 2 and 3 was driven by the perceived difficulty of the question, then participants in Experiment 3 will favor group reasoning more for the *Hard Percept* questions than the *Easy Reasoning* questions.

A.1.3 Cross-Experiment Exploratory Analyses

We conducted several exploratory analyses comparing results between experiments to examine the effects of crowd size and and question type more broadly. Experiment 1 and Experiment 2 used identical questions, but Experiment 2 increased the size of the crowd from 5 to 50 people. Our preregistered prediction was that participants would favor the crowd for population preference questions, but continue to favor the group for reasoning questions. However, we can also test the direct effect of crowd size by comparing people's judgments for reasoning and for popularity questions in Experiment 1 to their judgments in Experiment 2. Experiment 3 again used a crowd of 50 people, but contrasted easy versions of the reasoning questions from Experiments 1 and 2 with challenging perceptual discrimination tasks. This allowed us to test whether the preference for group discussion was caused by the perceived difficulty of the question. However, it also allows us to test whether the preference for crowdsourcing observed in

Experiments 1 and 2 extended to questions with a more ambiguous relationship to crowd size than population preferences.

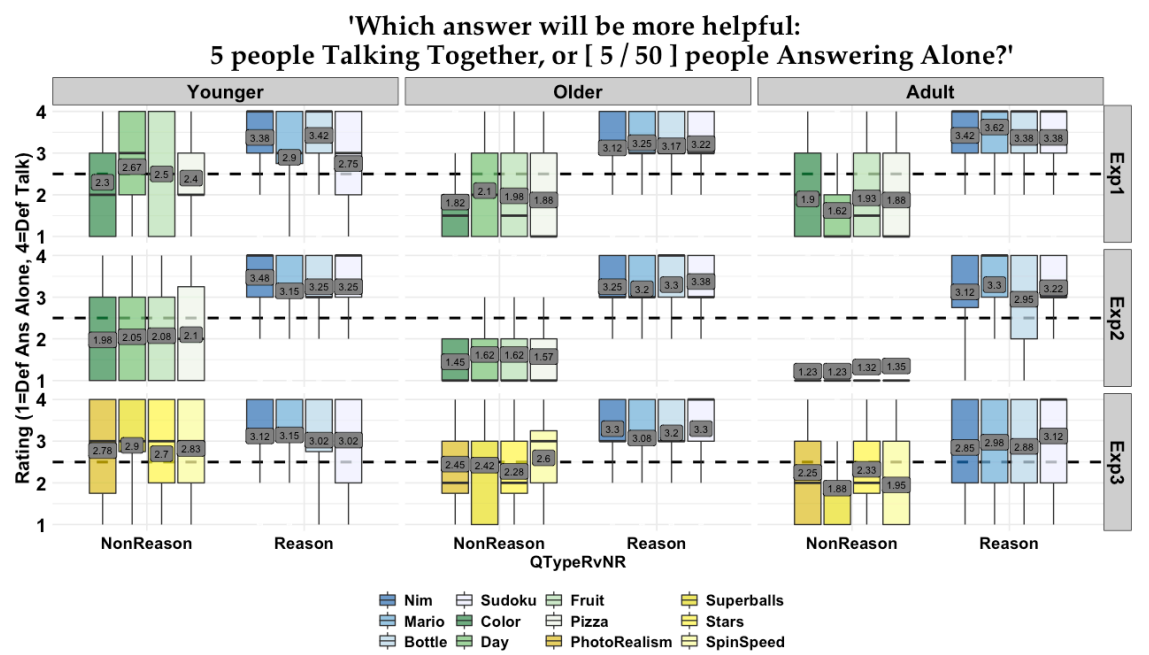
To explore the effect of crowd size, we ran separate ANOVAs for each QuestionType using AgeGroup & Experiment as predictors (Exps 1 and 2). The tenfold increase in crowd size had no impact on participants' preference for discussing reasoning questions in small groups ($F(1, 234)=0.045$, $p=.8320$); an *AgeGroup*ExpNum* interaction was significant ($F(2,234)=4.434$, $p=.0129$), but post-hoc comparisons revealed only a marginal difference between younger children's and adults' preference for reasoning in groups in Exp 1, but no other differences. However, participants were significantly more likely to crowdsource popularity questions in Experiment 2 than Experiment 1 ($F(1, 234)=19.303$, $p<0.0001$), with no differences between age groups.

To explore whether the crowdsourcing preference was as strong for perceptual discrimination problems as population preference questions, we ran an ANOVA comparing the two types of non-reasoning questions, using AgeGroup & Experiment as predictors (Exps 2 and 3). Participants were significantly less confident that crowdsourcing would be preferable to a small group discussion for percept questions than popularity questions ($F(1, 234)=76.897$, $p < 0.0001$); the interaction was not significant ($F(2,234)=0.139$, $p=.87$). Notably however, there was no difference between participants' preference for asking a small group to discuss *Easy Reasoning* questions in Experiment 3 and *Reasoning* questions in Experiment 2, though it did approach significance ($F(1, 234)=3.858$, $p < 0.0507$).

A.1.4 mTurk Quality Screen

We present instructions as voice-over videos in order to prevent language bots from skimming the written text, and immediately after the videos, we simply present participants with 3 multiple choice questions about their task (*A: is their job to answer the questions themselves or decide which answer will help Jack more, B: do the people who answer alone talk together before each telling Jack their answer or not talk together, C: do the people who talk together each tell Jack their own answer after talking, or do they have to agree on a single answer to tell Jack after talking*), with the correct answer being a nearly verbatim transcript from the video. Participants who get 1 or more of the attention check questions wrong have one more opportunity to answer after watching the video again; if they get any questions wrong in the second round, they're blocked from taking the survey.

A.1.5 Supplemental Plot for Exps 1-3



By-question boxplots for each of the 4 Reasoning questions (Blues) and 4 Non-Reasoning questions (Popularity - Greens; Percept - Yellows) in each experiment. Grey labels are means. (For preregistered analyses, average scores were computed for each QuestionType).

A.1.6 Mixed Effects Models

Our preregistered analysis plan was to compute an average score from the four questions of each QuestionType and conduct a repeated measure ANOVA on these two average scores. However, we also report mixed effects models; by including the un-averaged ratings for each question (i.e., the ratings on the 4-point scale for each of the four questions of each question type), these account for variance in the questions themselves. For each experiment, we tested the model (Ct_Rating ~ 0+AgeGroup*QuestionType + (1 | subID), which models the responses for each of the 8 questions while treating AgeGroup and QuestionType as fixed effects, and allowing random intercepts for each subject. Centering individual ratings on 2.5 and deleting the intercept compares simple effect estimates to "chance" (i.e., 2.5 on a scale of 1 to 4) for each age group and estimates of interactions to the prior level's interaction, testing our predictions versus chance for the reference level of QuestionType and versus the magnitude of the previous age group's interaction for each interaction term; we report models with both Reasoning and Non-Reasoning questions coded as the reference level. These MEMs of raw ratings for each question produced qualitatively identical results to the repeated measures ANOVA on the averaged question ratings, with one exception: in Experiment 3, the mixed effect model suggested that while the youngest children favored group discussion for Non-Reasoning questions as well as Reasoning questions (consistent with the ANOVA), they also distinguished between the two (contrary to the ANOVA, where the difference

was not significant), favoring discussion for Reasoning question more than for Non-Reasoning questions

(A) Exp 1: All age groups favored group discussion for Reasoning questions ($\beta_{Younger}$ = .6125, *SE* = .086, *p* = 9.33e-12; β_{Older} = .69375, *SE* = .086, *p* = 2.15e-14; β_{Adult} = .950, *SE* = .086, *p* < 2e-16), as well as making increasingly stronger distinctions between Reasoning and Non-Reasoning questions with age ($\beta_{Younger}$ = -.64375, *SE* = .10774, *p* = 3.40e-09; β_{Older} = -.60625, *SE* = .15236, *p* = 7.52e-05; β_{Adult} = .950, *SE* = .15236, *p* < 2.60e-10). Rerunning the regression with Non-Reasoning as the reference level showed that while younger children did not favor crowdsourcing for Non-Reasoning questions, older children and adults did ($\beta_{Younger}$ = -.03125, *SE* = .086, *p* = 0.717; β_{Older} = -.55625, *SE* = .086, *p* = 4.62e-10; β_{Adult} = -.66875, *SE* = .086, *p* < 1.47e-13)

(B) Exp 2: As in Exp 1, all age groups favored group discussion for Reasoning questions ($\beta_{Younger}$ = .78125, *SE* = .086, *p* < 2e-16; β_{Older} = .78125, *SE* = .086, *z p* < 2e-16; β_{Adult} = .65, *SE* = .086, *p* = 9.77e-13), as well as making increasingly stronger distinctions between Reasoning and Non-Reasoning questions with age ($\beta_{Younger}$ = -1.23125, *SE* = .092, *p* < 2e-16; β_{Older} = -.48125, *SE* = .130, *p* = 0.000232; β_{Adult} = -.63750, *SE* = .130, *p* =1.16e-06). Rerunning the regression with Non-Reasoning as the reference level showed that all age groups favored crowdsourcing for Non-Reasoning questions ($\beta_{Younger}$ = -.450, *SE* = .086, *p* = 3.73e-07; β_{Older} = -.93125, *SE* = .086, *p* < 2e-160; β_{Adult} = -.1.21875, *SE* = .086, *p* < 2e-16).

(C) Exp 3: All age groups favored group discussion for Reasoning questions ($\beta_{Younger}$ = .58125, *SE* = .096, *p* = 5.64e-09; β_{Older} = .71875, *SE* = .096, *p* = 1.45e-12; β_{Adult} = .45625, *SE* = .096, *p* = 3.56e-06), as well as making increasingly stronger distinctions between Reasoning and Non-Reasoning questions with age ($\beta_{Younger}$ = -.28125, *SE* = .106, *p* =0.008342; β_{Older} = -.500, *SE* = .150, *p* = 0.000926; β_{Adult} = -.575, *SE* = .150, *p* = 0.000142). Rerunning the regression with Non-Reasoning as the reference level showed that while younger children preferred to discuss Non-Reasoning questions as well, older children had no preference, and adults preferred crowdsourcing Non-Reasoning questions ($\beta_{Younger}$ = .300, *SE* = .096, *p* = 0.002019; β_{Older} = -.06250, *SE* = .096, *p* = 0.516040; β_{Adult} = -.400, *SE* = .096, *p* = 4.4e-05).

More complex random effects specification were overfit or failed to converge, but suggested little variance between questions themselves after accounting for the effect of QuestionType. For instance, Model 1 (below) allows for random intercepts of questions within the fixed effect of QuestionType, but fit was singular. Inspecting random effects suggested that the (1 | QuestionType:Question) explained no variance.

Model 1: Ct_Rating ~ 0+QuestionType*AgeGroup+(1 | subID)+(1 | QuestionType:Question)

A.2 Supplement to Chapter 3

A.2.1 Supplement Experiment 2: Methods & Results for Competence Judgments

Method. Past work has suggested that children think that "fast = smart" (Heyman & Compton, 2006); accordingly, younger children in particular could reason that "faster agents are smarter, and smarter agents are better at solving puzzles, so a faster response is more likely to be accurate". However, Heyman & Compton (2006) verbally described agents' response time and task difficulty to children, leaving children no opportunity to decide how fast was too fast. When allowed to evaluate time and difficulty for themselves, children might show more skepticism. Thus, in order to compare our results to past work, we asked children to make a competence judgment at the end of the experiment. Children were shown two new figures and told that these agents would "each start their engines when they think they've figured out the shortest road that follows the rules". One agent started his engine at 3s, and the other at 20s. However, unlike in main task, we then told the children that both agents had answered correctly. Children were then asked which of the two agents was "better at this game". If children have the "fast = better" bias observed in prior studies, then of two agents who both accurately solve a problem, children should believe that the faster agent is more competent.

Results. For the Ability question , younger children and adults did not differ from chance when asked whether the fast or slow agent was "better" at the game, but older children were more likely to choose the faster agent (Supplemental Fig. 1: $M_{Younger}$=53.3%, binomial p=*ns*, M_{Older}=71.1%, binomial p=.007, binomial M_{Adult}=60.0%, p = *ns*). Past research using vignettes about agents solving puzzles has suggested that while children have a "faster=better" bias, children under 7 also confound effort, ability, and outcome, while older children begin to attribute outcomes to some combination of effort and ability (Heyman & Compton, 2006; Nicholls, 1978; Stipek & Iver, 1989). Children's judgments in the second task were consistent with that developmental trajectory: when told that the outcome was the same for the fast and slow agent, the 5-7 year olds considered them equally competent, while older children appeared to infer that the difference in speed nevertheless implied a difference in competence. However, adults did not consider the faster agent more competent in the Ability task. Adults may have been rightly been skeptical of the fast agent's competence advantage. Puzzles as complex as the ones we used could only be "solved" in three seconds by a lucky guess.

A.2.2 Supplement Experiment 2: Comparing responses on the 1st and 2nd task

If an agent's response speed triggers a joint inference about their competence, this could offer a competing account of children's accuracy judgments in the main task of Experiment 2. Comparing children's competence judgments in the 2nd task with their accuracy judgments in the first task offer could some evidence for or against this account. If participants thought that the slow agent in the main task was less competent than the fast agent, there are three patterns we might expect to observe in the data. First, they might expect the fast agent to be more accurate than the slow agent; however, our primary results directly contradict that prediction — indeed,

the reverse is true. Second, participants who judged the fast agent as more competent than the slow agent in the 2nd task may also have expected the slow agent's solution to be less accurate (in an absolute sense) in the main task. However, we found no association between participants' average accuracy predictions for the slow agent in the first task and their relative competence judgments in the second task, for any age group or the sample as a whole (Models 1a-1c in Table 1 below). Third, we might expect an association between the *difference* in average accuracy predictions for the fast and slow agents on the first task and children's competence judgments on the 2nd task. However, we again found no association here for any age group or the sample as a whole (Models 2a-2c in Table 2 below).

Table 1

term	estimate	std.error	statistic	p.value
Mod1a				
(Intercept)	2.56	0.10	24.82	0.00
scale(AgeYears)	0.13	0.10	1.22	0.23
Ability_Num	0.03	0.13	0.22	0.83
scale(AgeYears):Ability_Num	-0.10	0.13	-0.73	0.47
Mod1b				
(Intercept)	2.40	0.13	18.66	0.00
AgeGroupOlder	0.37	0.21	1.79	0.08
AgeGroupAdult	0.34	0.19	1.82	0.07
Ability_Num	0.17	0.18	0.98	0.33
AgeGroupOlder:Ability_Num	-0.33	0.26	-1.25	0.21
AgeGroupAdult:Ability_Num	-0.13	0.25	-0.53	0.60
Mod1c				
(Intercept)	2.61	0.08	31.67	0.00
Ability_Num	0.05	0.11	0.44	0.66

Table 2

term	estimate	std.error	statistic	p.value
Mod2a				
(Intercept)	0.56	0.25	2.24	0.03
scale(AgeYears)	0.32	0.26	1.23	0.22
SF_Diff	-0.14	0.28	-0.51	0.61
scale(AgeYears):SF_Diff	0.02	0.26	0.08	0.94
Mod2b				
(Intercept)	0.17	0.31	0.54	0.59
AgeGroupOlder	0.80	0.50	1.61	0.11
AgeGroupAdult	-0.14	0.58	-0.24	0.81
SF_Diff	-0.14	0.37	-0.39	0.70
AgeGroupOlder:SF_Diff	-0.01	0.55	-0.01	0.99
AgeGroupAdult:SF_Diff	0.53	0.54	0.98	0.33
Mod2c				
(Intercept)	0.44	0.21	2.10	0.04
SF_Diff	0.04	0.21	0.20	0.84

A.2.3 Supplement Experiment 3: Change to Materials

Materials. We made two changes to the materials in Experiment 3. In addition to generating a set of Easy maps, we used live videos of agents drawing one of two cards from a stack (instead of having a single map appearing on a screen before an animated 2D silhouette), looking at it, and ringing a bell to signal completion after either 3s or 20s. This was done in order to clarify what the agent had visual access to and when, while simultaneously showing children the two possible maps on the powerpoint slide in which the video was embedded. Each map appeared on either a blue or green background, so that children could answer by simply saying a color. All materials, along with information about the videos and counterbalancing, can be found on the OSF repository.

A.2.4 Supplement Experiment 3: Results for Second Task ("Which Map", when difficulty equalized)

In a second task at the end of the experiment, participants saw two additional trials, in which an agent drew a card with one of two *Easy* puzzles or one of two *Hard* puzzles and claimed to have solved the puzzle after 3s. Participants were then asked which map the agent had been looking at. Deprived of task difficulty as a cue, participants were no more likely to infer that the agent had drawn one than the other (*Age6*: M_{Easy}=46.9%, p=0.860., M_{Hard}=53.1%, p=0.860; *Age7*: M_{Easy}= 59.4%, p=0.377, M_{Hard}=46.9%, p=0.860; *Age8*: M_{Easy}= 46.9%, p=0.860, M_{Hard}= 56.2%, p=0.597; *Adults*: M_{Easy}= 32.3%, p=0.071, M_{Hard}= 50.0%, p=1.00).

A.2.5 Supplemental Analysis: Ordinal Regressions

The response variable for the main task in all three experiments was a 1-4 ordinal rating scale. Though we preregistered analyses based on standard regression techniques, which treat ordinal scales as metric data, some recent publications strongly recommend that ordinal data never be treated as metric due to the increased risk of both Type I and Type II errors (Liddell & Kruschke, 2018) compared to alternative analyses. Thus, we also also include an alternative analysis below which uses cumulative ordinal regressions with logit link functions. Instead of assuming that the distance between each level of the response to an ordinal scale (e.g., "definitely accurate" vs. "probably accurate") is equal, ordinal models assume that responses are mapped on to the ordered set of scale categories from an underlying continuous distribution, and search for the latent thresholds that "cut" the continuous distribution into response categories. Intuitively, a ordinal regression can be thought of a set of logistic regressions moving in order from one cut threshold to the next. Thus, in addition to producing the odds ratios for each coefficient in the model, an ordinal regression produces *k-1* thresholds, where *k* is the number of categories in the ordinal (e.g., Likert) scale. Importantly, while the standard coefficients reflect change across levels, the threshold cuts are the same for all levels of all variables. However, just as in logistic regression models, the thresholds and standard coefficients can be exponentiated to produce predicted probabilities.

We used the **clmm** function from the **ordinal** package in R, which has similar random effects syntax to the lme4 package. We conducted two kinds of ordinal regression for each Experiment, including by-participant random intercepts in each. Model 1 looks for between condition differences; when Model 2 is run on each Age and RTSpeed separately, exponentiating the threshold cut between levels 2 and 3 of the 4-point Likert scale compares ratings to chance.

(Model 1): clmm(as.ordered(Ratings) ~ RTSpeed*AgeGroup + (1|subID), data=., link="logit")

(Model 2): clmm(as.ordered(Ratings) ~ 1 + (1|subID), data=., link="logit")

Exp 1: Results of Ordinal Regression. The results of the ordinal regression were similar to those of the ANOVA. Model 1 suggested that the age groups as a whole were more likely to infer that slow responders were figuring the maps out for first time than fast responders (Log-

$OR_{SpeedSlow}$=1.16, *SE* = *.23*, *z* = 5.01, *p*=.001). Moreover, both older children and adults were more likely than younger children to infer that the fast responders were remembering than figuring out (Log-$OR_{AgeOlder}$=-0.55, *SE* = .23, *z* = -2.39, *p*=.017; Log-$OR_{AgeAdult}$=-1.20, *SE* = .24, *z* = -4.99, *p*=.001), with the difference between fast and slow agents also increasing across *AgeGroups* (Log-$OR_{Older*Slow}$ = 1.35, *SE* = .32, *z* = 4.20, *p*=.001; Log-$OR_{Adult*Slow}$ = 2.52, *SE* = .34, *z* = 7.51, *p*=.001). Exponentiating the 2|3 threshold coefficient from Model 2 produced similar results, suggesting that all age groups also inferred that Fast responders were more likely to be remembering answers than figuring out: the model predicts that 65.7% of younger children (*z* = 2.82, *95CI:* 54.9—75.0), 83.2% of older children (*z* = 5.33, *95CI:* 73.3—89.9), and 95.5% of adults (*z* = 5.92, *95CI:* 88.5—98.3) would give a rating of 2 or less on the 4-point scale. The reverse was true for Slow responders: the model predicts that 67.3% of younger children (*z* = -3.58, *95CI:* 58.1–75.3), 84.3% of older children (*z* = -4.84, *95CI:* 73.1—91.39), and 90.9% of adults (*z* = -5.80, *95CI:* 82.1—95.6) would give a rating of 3 or *more* on the 4-point scale. Thus, the only difference between the ANOVA and ordinal regression was the inclusion of coefficients estimating the strength of the developmental shift across age groups.

Exp 2: Results of Ordinal Regression. The results of the ordinal regression were similar to those of the ANOVA. Model 1 suggested that all age groups were more likely to infer that slow responders were accurate than inaccurate (Log-$OR_{SpeedSlow}$=0.53, *SE* = .23, *z* = 2.25, *p*=.024). Additionally, adults were more likely than younger children to infer that the fast responders were inaccurate than accurate, though older children did not differ from younger children (Log-$OR_{AgeOlder}$=-0.55, *SE* = .23, *z* = -2.39, *p*=.017; Log-$OR_{AgeAdult}$=-1.20, *SE* = .24, *z* = -4.99, *p*=.001); similarly, while the difference between inferences for fast and slow ages was significantly greater for adults than younger children, the difference was no greater in older children than younger children (Log-$OR_{Older*Slow}$ = 1.35, *SE* = .32, *z* = 4.20, *p*=.001; Log-$OR_{Adult*Slow}$ = 2.52, *SE* = .34, *z* = 7.51, *p*=.001). Exponentiating the 2|3 threshold coefficient from Model 2 suggested that all age groups also inferred that Fast responders were more likely be inaccurate than accurate: Model 2 predicts that 62.2% of younger children (*z* = 2.81, *95CI:* 54.9—70.0), 64.3% of older children (*z* = 2.98, *95CI:* 55.0—72.6), and 86.9% of adults (*z* = 6.39, *95CI:* 78.8—92.2) would give a rating of 2 or less on the 4-point scale. However, while Model 2 predicts that 74.3% of adults (*z* = 2.89, *95CI:* 58.4—8.56) inferred that Slow responders were more likely accurate than inaccurate, the difference for younger children was not significant (48.1%, *z* = 0.43, *95CI:* 39.4—56.9). For older children, the model including random intercepts by-participant was singular; when the by-participant intercepts were removed, the model suggested that 70.9% of responses from older children rated Slow responders as more likely accurate than inaccurate, a significant difference (*z* = 1.97, *p* = .0489). Thus, as in Experiment 1, the only difference between the ANOVA and ordinal regression was the inclusion of coefficients estimating the strength of the developmental shift across age groups.

Exp 3: Results of Ordinal Regression. The results of the ordinal regression were similar to those of the ANOVA. We first asked whether the child sample would be more likely to infer that fast responses were easy maps than hard maps, using a model with random intercepts by participant, and age (centered on the mean of the child sample (7)) as a fixed effect. The effect of Ct_Age was

significant (Log-OR_{Ct_Age}=-0.48, *SE* = *.24*, *z* = -2.05, *p*=.041). Exponentiating the 2|3 threshold coefficient of Model 2 suggested that 82.9% of 6-year-olds (*z* = 2.88, *95CI:* 62.4—93.4), 78.3% of 7-year-olds (*z* = 3.25, *95CI:* 62.4—88.6), and 91.8% of 8-year-olds (*z* = 97.0, *95CI:* 79.5—97.0) would infer the fast agents to be solving easy maps rather than hard maps. Similarly, when both Ct_Age and RTSpeed were included as fixed effects, participants inferred that fast agents were more likely than slow agents to be solving easy maps (Log-$OR_{SpeedSlow}$=1.27, *SE* = .20, *z* = 6.23, *p* <.001), with a significant effect of age and a marginal interaction (Log-OR_{Ct_Age}=-0.40, *SE* = .19, *z* = -2.05, *p*=.04; Log-$OR_{Age*Speed}$=0.47, *SE* = .24, *z* = 1.96, *p*=.05). While neither the age effect nor the interaction reached significance in the analogous ANOVA presented in the main paper, these effects do not change the results: running a separate model of RTSpeed for each age as well as adults and exponentiating the 2|3 threshold coefficient to produce odds ratios suggested that the difference between fast and slow trials was significant for all age groups ($OR_{SlowAge6}$= 2.26, *SE* = .33, *z* = 2.45, *p* = .014; $OR_{SlowAge7}$ = 3.24, *SE* = .35, *z* = 3.38, *p* < .001; $OR_{SlowAge8}$ = 6.54, *SE* = .40, *z* = 4.75, *p* < .001; $OR_{SlowAdult}$ = 12.10, *SE* = 1.77, *z* = 4.86, *p* < .001). In the second task of Experiment 3, participants saw two trials in which an agent quickly responded to one of two hard maps or one of two easy maps and participants were asked to rate their accuracy on a 4-point scale after inferring which map was the agent's target (with no difference in the difficulty of the two maps, participants were equally likely to infer either as the agents' target; see main text). The fit of a model including random intercepts by-participant was singular, so we removed them; in the reduced model, a significant effect of difficulty level suggested that participants were more likely to infer that the agent's solution was accurate for the easy map than the hard map (Log-OR_{Item_Hard} = -1.28, *SE* = .49, *z* = -2.64, *p*=.008), with the inferred accuracy of the easy map increasing with age (Log-OR_{Age7} = 1.10, *SE* = .47, *z* = 2.35, *p*=.019; Log-OR_{Age8} = 1.01, *SE* = .48, *z* = 2.10, *p*=.036; Log-OR_{Adult} = 2.30, *SE* = .51, *z* = 4.51, *p* < .001). There was a significant interaction of age and difficulty for adults, but not for any other age group (Log-$OR_{Adult*Item\_Hard}$ = -1.46, *SE* = .68, *z* = -2.12, *p*=.034). Running a separate model for the easy and hard maps for each age and exponentiating the 2|3 threshold coefficient to produce odds ratios suggested that, as with the ANOVA analysis in the main paper, absolute estimations of accuracy were less clear. Children ages 6 and 8, but not adults or 7-year-olds, believed that the agent's solution was inaccurate for the *Hard* puzzle (Log-OR_{Age6} = 3.00, *SE* = .41, *z* = 2.69, *p* = .007; Log-OR_{Age7} = 1.91, *SE* = .37, *z* = 1.74, *p* = .082; Log-OR_{Age8} = 5.40, *SE* = .49, *z* = 3.46, *p* < .001; Log-OR_{Adult} = 1.58, *SE* = .37, *z* = 1.25, *p* = .213), while children ages 7 and 8, but not age 6, believed that the agent's solution was accurate for the *Easy* puzzle (Log-OR_{Age6} = 0.88, *SE* = .35, *z* = -0.35, *p* = .724; Log-OR_{Age7} = 0.28, *SE* = .43, *z* = -2.98, *p* = .003; Log-OR_{Age8} = 0.39, *SE* = .39, *z* = -2.39, *p* = .017). Similarly, no adult rated accuracy on the Easy maps as less than 3 ("probably accurate") on the 4-point scale; this high confidence in accuracy on the Easy map prevented the cumulative link model from estimate a 2|3 threshold. In sum, as in Experiments 1 and 2, the only difference between the ANOVA and ordinal regression was the inclusion of coefficients estimating the strength of the developmental shift across age groups.

References

Abel, M., & Bäuml, K.-H. T. (2020). Social interactions can simultaneously enhance and distort memories: Evidence from a collaborative recognition task. *Cognition*, *200*, 104254. https://doi.org/10.1016/j.cognition.2020.104254

Abney, D. H., Suanda, S. H., Smith, L. B., & Yu, C. (2020). What are the building blocks of parent–infant coordinated attention in free-flowing interaction? *Infancy*, *25*(6), 871–887. https://doi.org/10.1111/infa.12365

Aboody, R., Davis, Isaac, Dunham, Y., & Jara-Ettinger, J. (2021). I can tell you know a lot, although I'm not sure what: Modeling broad epistemic inference from minimal action. *Proceedings of the Cognitive Science Society*, 6. https://doi.org/10.31234/osf.io/uymtz

Aboody, R., Denison, S., & Jara-Ettinger, J. (2021, May 11). Children consider the probability of random success when evaluating knowledge. *Proceedings of the Cognitive Science Society*. https://doi.org/10.31234/osf.io/a7g9t

Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. C. (2019). Says who? Children consider informants' sources when deciding whom to believe. Cognitive Development Society, Louisville, KY.

Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. C. (2022). Says who? Children consider informants' sources when deciding whom to believe. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001198

Aboody, R., Zhou, C., & Jara-Ettinger, J. (2021). In Pursuit of Knowledge: Preschoolers Expect Agents to Weigh Information Gain and Information Cost When Deciding Whether to Explore. *Child Development*, *92*(5), 1919–1931. https://doi.org/10.1111/cdev.13557

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, *14*(1), 43–64. https://doi.org/10.1080/09515080120033571

Albrecht, J., Anderson, A., & Vroman, S. (2010). Search by committee. *Journal of Economic Theory*, *145*(4), 1386–1407. https://doi.org/10.1016/j.jet.2009.05.011

Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences*, *118*(36), e2101062118. https://doi.org/10.1073/pnas.2101062118

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 201917687. https://doi.org/10.1073/pnas.1917687117

Almaatouq, A., Rahimian, M. A., Burton, J. W., & Alhajri, A. (2020). *When social influence promotes the wisdom of crowds*. PsyArXiv.

Anderson, L. R., & Holt, C. A. (1997). Information Cascades in the Laboratory. *The American Economic Review*, *87*(5), 17. https://www.jstor.org/stable/2951328

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914–930. https://doi.org/10.1037/0278-7393.33.5.914

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1350–1365. https://doi.org/10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4). https://doi.org/10.1038/s41562-017-0064

Barkoczi, D., & Galesic, M. (2016). Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, *7*(1). https://doi.org/10.1038/ncomms13109

Becker, J., Almaatouq, A., & Horvat, E. A. (2020). Network Structures of Collective Intelligence: The Contingent Benefits of Group Discussion. *PsyArxiv*.

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 201615978. https://doi.org/10.1073/pnas.1615978114

Becker, J., Porter, E., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, *116*(22), 10717–10722. https://doi.org/10.1073/pnas.1817195116

Berke, M., & Jara-Ettinger, J. (2021, May 20). Thinking about thinking through inverse reasoning. *Proceedings of the Cognitive Science Society*. https://doi.org/10.31234/osf.io/r25qn

Boehm, C. (1996). Emergency Decisions, Cultural-Selection Mechanics, and Group Selection [and Comments and Reply]. *Current Anthropology*, *37*(5), 763–793. https://doi.org/10.1086/204561

Bonner, B. L., Shannahan, D., Bain, K., Coll, K., & Meikle, N. L. (2021). The Theory and Measurement of Expertise-Based Problem Solving in Organizational Teams: Revisiting Demonstrability. *Organization Science*, orsc.2021.1481. https://doi.org/10.1287/orsc.2021.1481

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding How To Decide: Self-Control and Meta-Decision Making. *Trends in Cognitive Sciences*, *19*(11), 700–710. https://doi.org/10.1016/j.tics.2015.08.013

Boyd, R., & Richerson, P. J. (1995). Boyd R, Richerson PJ (1995) Why does culture increase human adaptability. *Ethology & Sociobiology*, *16*, 125–143. https://doi.org/10.1016/0162-3095(94)00073-G

Brandon, D. P., & Hollingshead, A. B. (2004). Transactive Memory Systems in Organizations: Matching Tasks, Expertise, and People. *Organization Science*, *15*(6), 633–644. https://doi.org/10.1287/orsc.1040.0069

Brennan, S., E., & Williams, M. (1995). The feeling of Another's Knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*, 383–398. https://doi.org/10.1006/jmla.1995.1017

Bruce, R. S. (1935). Group Judgments in the Fields of Lifted Weights and Visual Discrimination. *The Journal of Psychology*, *1*(1), 117–121. https://doi.org/10.1080/00223980.1935.9917245

Burdett, E. R. R., Lucas, A. J., Buchsbaum, D., McGuigan, N., Wood, L. A., & Whiten, A. (2016). Do Children Copy an Expert or a Majority? Examining Selective Learning in Instrumental and Normative Contexts. *PLOS ONE*, *11*(10), e0164698. https://doi.org/10.1371/journal.pone.0164698

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, *359*(6381), 1263–1266. https://doi.org/10.1126/science.aao3539

Centola, D. (2022). The network science of collective intelligence. *Trends in Cognitive Sciences*, S1364661322002054. https://doi.org/10.1016/j.tics.2022.08.009

Chan, J., Lizzeri, A., Suen, W., & Yariv, L. (2018). Deliberating Collective Decisions. *The Review of Economic Studies*, *85*(2), 929–963. https://doi.org/10.1093/restud/rdx028

Chi, M., Roy, M., & Hausmann, R. (2008). Observing Tutorial Dialogues Collaboratively: Insights About Human Tutoring Effectiveness From Vicarious Learning. *Cognitive Science: A Multidisciplinary Journal*, *32*(2), 301–341. https://doi.org/10.1080/03640210701863396

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting Self-Explanations Improves Understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1207/s15516709cog1803_3

Chittka, L., Skorupski, P., & Raine, N. E. (2009). Speed–accuracy tradeoffs in animal decision making. *Trends in Ecology & Evolution*, *24*(7), 400–407. https://doi.org/10.1016/j.tree.2009.02.010

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. http://www.jstor.org/stable/3328150

Clark, H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Conradt, L., & List, C. (2009). Group decisions in humans and animals: A survey. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 719–742. https://doi.org/10.1098/rstb.2008.0276

Cooney, G., Mastroianni, A. M., Abi-Esber, N., & Brooks, A. W. (2020). The many minds problem: Disclosure in dyadic versus group conversation. *Current Opinion in Psychology*, *31*, 22–27. https://doi.org/10.1016/j.copsyc.2019.06.032

Cottrell, S., Torres, E., Harris, P. L., & Ronfard, S. (2023). Older children verify adult claims because they are skeptical of those claims. *Child Development*, *94*(1), 172–186. https://doi.org/10.1111/cdev.13847

Couzin, I. D., Ioannou, C. C., Demirel, G., Gross, T., Torney, C. J., Hartnett, A., Conradt, L., Levin, S. A., & Leonard, N. E. (2011). Democratic Consensus in Animal Groups. *Science*, *334*(6062), 4. https://doi.org/10.1126/science.1210280

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133. https://doi.org/10.1207/s15516709cog2701_4

Danovitch, J. H., & Keil, F. C. (2007). Choosing between hearts and minds: Children's understanding of moral advisors. *Cognitive Development*, *22*(1), 110–123. https://doi.org/10.1016/j.cogdev.2006.07.001

de Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences*, *115*(9), 2066–2071. https://doi.org/10.1073/pnas.1717632115

Derex, M., Beugin, M.-P., Godelle, B., & Raymond, M. (2013). Experimental evidence for the influence of group size on cultural complexity. *Nature*, *503*(7476), 389–391. https://doi.org/10.1038/nature12774

Derex, M., Bonnefon, J.-F., Boyd, R., & Mesoudi, A. (2019). Causal understanding is not necessary for the improvement of culturally evolving technology. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0567-9

Derex, M., & Boyd, R. (2015). The foundations of the human cultural niche. *Nature Communications*, *6*(1). https://doi.org/10.1038/ncomms9398

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, *113*(11), 2982–2987. https://doi.org/10.1073/pnas.1518798113

Derex, M., Perreault, C., & Boyd, R. (2018). Divide and conquer: Intermediate levels of population fragmentation maximize cultural accumulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1743), 20170062. https://doi.org/10.1098/rstb.2017.0062

Dietrich, F., & Spiekermann, K. (2013a). Epistemic Democracy with Defensible Premises. *Economics and Philosophy*, *29*, 34. https://doi.org/doi:10.1017/S0266267113000096

Dietrich, F., & Spiekermann, K. (2013b). Independent Opinions? On the Causal Foundations of Belief Formation and Jury Theorems. *Mind*, *122*(487), 655–685. https://doi.org/10.1093/mind/fzt074

Domberg, A., Köymen, B., & Tomasello, M. (2019). Children choose to reason with partners who submit to reason. *Cognitive Development*, *52*, 100824. https://doi.org/10.1016/j.cogdev.2019.100824

Droit-Volet, S., & Wearden, J. H. (2001). Temporal Bisection in Children. *Journal of Experimental Child Psychology*, *80*(2), 142–159. https://doi.org/10.1006/jecp.2001.2631

Dunn, J. (2019). Reliable group belief. *Synthese*. https://doi.org/10.1007/s11229-018-02075-8

Novaes, C. D. (2020). The Dialogical Roots of Deduction: Historical, Cognitive, and Philosophical Perspectives on Reasoning. Cambridge University Press.

Einav, S. (2014). Does the Majority Always Know Best? Young Children's Flexible Trust in Majority Opinion. *PLoS ONE*, *9*(8), e104585. https://doi.org/10.1371/journal.pone.0104585

Einav, S. (2018). Thinking for themselves? The effect of informant independence on children's endorsement of testimony from a consensus. *Social Development*, *27*(1), 73–86. https://doi.org/10.1111/sode.12264

Feghhi, I., & Rosenbaum, D. A. (2019). Judging the subjective difficulty of different kinds of tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(8), 983–994. https://doi.org/10.1037/xhp0000653

Feldman, J. A., & Ballard, D. H. (1982). Connectionist Models and Their Properties. *Cognitive Science*, *6*(3), 205–254. https://doi.org/10.1207/s15516709cog0603_1

Galton, F. (1907). Vox Populi. *Nature*, *75*, 450–451. https://doi.org/doi.org/10.1038/075450a0

Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. L. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition*, *217*, 104885. https://doi.org/10.1016/j.cognition.2021.104885

Goldman, A. I. (2014). Social Process Reliabilism. In J. Lackey (Ed.), *Essays in Collective Epistemology* (pp. 11–41). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199665792.003.0002

Goldstone, R. L., & Theiner, G. (2017). The multiple, interacting levels of cognitive systems (MILCS) perspective on group cognition. *Philosophical Psychology*, *30*(3), 338–372. https://doi.org/10.1080/09515089.2017.1295635

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461–482. https://doi.org/10.1037/0033-295X.113.3.461

Griggs, R. A. (2015). The Disappearance of Independence in Textbook Coverage of Asch's Social Pressure Experiments. *Teaching of Psychology*, *42*(2), 137–142. https://doi.org/10.1177/0098628315569939

Grocke, P., Rossano, F., & Tomasello, M. (2018). Young children are more willing to accept group decisions in which they have had a voice. *Journal of Experimental Child Psychology*, *166*, 67–78. https://doi.org/10.1016/j.jecp.2017.08.003

Grueneisen, S., & Tomasello, M. (2019). Children use rules to coordinate in a social dilemma. *Journal of Experimental Child Psychology*, *179*, 362–374. https://doi.org/10.1016/j.jecp.2018.11.001

Guarnaschelli, S., McKelvey, R. D., & Palfrey, T. R. (2000). An Experimental Study of Jury Decision Rules. *American Political Science Review*, *94*(2), 407–423. https://doi.org/10.2307/2586020

Gummerum, M., Leman, P. J., & Hollins, T. S. (2014). How do children share information in groups? *Developmental Psychology*, *50*(8), 2105–2114. https://doi.org/10.1037/a0037144

Gunn, L. J., Chapeau-Blondeau, F., McDonnell, M. D., Davis, B. R., Allison, A., & Abbott, D. (2016). Too good to be true: When overwhelming evidence fails to convince. *Proceedings of the*

Royal Society A: Mathematical, Physical and Engineering Sciences, *472*(2187), 20150748. https://doi.org/10.1098/rspa.2015.0748

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks. *Proceedings of the Cognitive Science Society*, 6.

Hahn, U., von Sydow, M., & Merdes, C. (2019). How Communication Can Make Voters Choose Less Well. *Topics in Cognitive Science*, *11*(1), 194–206. https://doi.org/10.1111/tops.12401

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

Hanus, D., Mendes, N., Tennie, C., & Call, J. (2011). Comparing the Performances of Apes (Gorilla gorilla, Pan troglodytes, Pongo pygmaeus) and Human Children (Homo sapiens) in the Floating Peanut Task. *PLoS ONE*, *6*(6), e19555. https://doi.org/10.1371/journal.pone.0019555

Hardwig, J. (1991). The Role of Trust in Knowledge. *The Journal of Philosophy*, *88*(12), 693–708. https://doi.org/10.2307/2027007

Harris, C. B., Keil, P. G., Sutton, J., Barnier, A. J., & McIlwain, D. J. F. (2011). We Remember, We Forget: Collaborative Remembering in Older Couples. *Discourse Processes*, *48*(4), 267–303. https://doi.org/10.1080/0163853X.2010.541854

Harris, P. L. (2002). What do children learn from testimony? In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The Cognitive Basis of Science* (pp. 316–334). Cambridge University Press. https://doi.org/10.1017/CBO9780511613517.018

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive Foundations of Learning from Testimony. *Annual Review of Psychology*, *69*(1), 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Hastie, R., & Kameda, T. (2005). The Robust Beauty of Majority Rules in Group Decisions. *Psychological Review*, *112*(2), 494–508. https://doi.org/10.1037/0033-295X.112.2.494

Haun, D. B. M., & Tomasello, M. (2011). Conformity to Peer Pressure in Preschool Children: Peer Pressure in Preschool Children. *Child Development*, *82*(6), 1759–1767. https://doi.org/10.1111/j.1467-8624.2011.01666.x

Haun, D. B. M., van Leeuwen, E. J. C., & Edelson, M. G. (2013). Majority influence in children and other animals. *Developmental Cognitive Neuroscience*, *3*, 61–71. https://doi.org/10.1016/j.dcn.2012.09.003

Heck, I. A., Bas, J., & Kinzler, K. D. (2021). Small groups lead, big groups control: Perceptions of numerical group size, power, and status across development. *Child Development*, *93*(1), 194-208. https://doi.org/10.1111/cdev.13670

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, *14*(3), 577–598. https://doi.org/10.1007/s11097-014-9355-1

Helwig, C. C., & Kim, S. (1999). Children's Evaluations of Decision-Making Procedures in Peer, Family, and School Contexts. *Child Development*, *70*(2), 502–512. https://doi.org/10.1111/1467-8624.00036

Heyes, C. (2016). Who Knows? Metacognitive Social Learning Strategies. *Trends in Cognitive Sciences*, *20*(3), 204–213. https://doi.org/10.1016/j.tics.2015.12.007

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press.

Heyes, C. (2019). Précis of Cognitive Gadgets: The Cultural Evolution of Thinking. *Behavioral and Brain Sciences*, *42*. https://doi.org/10.1017/S0140525X18002145

Heyman, G. D., & Compton, B. J. (2006). Context sensitivity in children's reasoning about ability across the elementary school years. *Developmental Science*, *9*(6), 616–627. https://doi.org/10.1111/j.1467-7687.2006.00540.x

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, *121*(1), 43–64. https://doi.org/10.1037/0033-2909.121.1.43

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Hu, J., Whalen, A., Buchsbaum, D., Griffiths, T., & Xu, F. (2015). Can Children Balance the Size of a Majority with the Quality of their Information? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 6.

Jackson, F. (1998). Reference and Description Revisited. *Nous*, *32*(S12), 201–218. https://doi.org/10.1111/0029-4624.32.s12.9

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Juni, M. Z., & Eckstein, M. P. (2017). The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences*, *114*(21), E4306–E4315. https://doi.org/10.1073/pnas.1610732114

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-022-00054-y

Kao, A. B., & Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1784), 20133305–20133305. https://doi.org/10.1098/rspb.2013.3305

Kao, A. B., Miller, N., Torney, C., Hartnett, A., & Couzin, I. D. (2014). Collective Learning and Optimal Consensus Decisions in Social Animal Groups. *PLoS Computational Biology*, *10*(8), e1003762. https://doi.org/10.1371/journal.pcbi.1003762

Keil, F. (2006). 6. Doubt, Deference, and Deliberation: Understanding and Using the Division of Cognitive Labor. In T. S. Gendler & J. P. Hawthorne (Eds.), *Oxford Studies in Epistemology* (Vol. 1, p. 24). Oxford University Press.

Keil, F. (2011). The Hidden Strengths of Weak Theories. *Anthropology and Philosophy*, *10*(1–2), 61. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4190847/

Keil, F. C. (2006). Explanation and Understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Kempe, M., & Mesoudi, A. (2014). An experimental demonstration of the effect of group size on cultural accumulation. *Evolution and Human Behavior*, *35*(4), 285–290. https://doi.org/10.1016/j.evolhumbehav.2014.02.009

Kerr, N. L., & Tindale, R. S. (2004). Group Performance and Decision Making. *Annual Review of Psychology*, *55*(1), 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

Kidd, C., White, K. S., & Aslin, R. N. (2011a). Learning the Meaning of "Um": Toddlers' developing use of speech disfluencies as cues to. In I. Arnon & E. V. Clark (Eds.), *Experience, Variation, and Generalization: Learning a First Language* (p. 28).

Kidd, C., White, K. S., & Aslin, R. N. (2011b). Toddlers use speech disfluencies to predict speakers' referential intentions: Toddlers use disfluencies to predict referential intentions. *Developmental Science*, *14*(4), 925–934. https://doi.org/10.1111/j.1467-7687.2011.01049.x

Kim, S., & Spelke, E. S. (2020). Learning from multiple informants: Children's response to epistemic bases for consensus judgments. *Journal of Experimental Child Psychology*, *192*, 104759. https://doi.org/10.1016/j.jecp.2019.104759

Kirschner, F., Paas, F., & Kirschner, P. A. (2009a). A Cognitive Load Approach to Collaborative Learning: United Brains for Complex Tasks. *Educational Psychology Review*, *21*(1), 31–42. https://doi.org/10.1007/s10648-008-9095-2

Kirschner, F., Paas, F., & Kirschner, P. A. (2009b). Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. *Computers in Human Behavior*, *25*(2), 306–314. https://doi.org/10.1016/j.chb.2008.12.008

Kitcher, P. (1990). The Division of Cognitive Labor. *The Journal of Philosophy*, *87*(1), 5. https://doi.org/10.2307/2026796

Kitcher, P. (1993). Knowledge, Society, and History. *Canadian Journal of Philosophy*, *23*(2), 155–177. https://doi.org/10.1080/00455091.1993.10717315

Klein, N., & Epley, N. (2015). Group discussion improves lie detection. *Proceedings of the National Academy of Sciences*, *112*(24), 7460–7465. https://doi.org/10.1073/pnas.1504048112

Kloo, D., Rohwer, M., & Perner, J. (2017). Direct and indirect admission of ignorance by children. *Journal of Experimental Child Psychology*, *159*, 279–295. https://doi.org/10.1016/j.jecp.2017.02.014

Knox, D., & Mummolo, J. (2020). Making inferences about racial disparities in police violence. *Proceedings of the National Academy of Sciences*, *117*(3), 1261–1262. https://doi.org/10.1073/pnas.1919418117

Kominsky, J. F., & Keil, F. C. (2014). Overestimation of Knowledge About Word Meanings: The "Misplaced Meaning" Effect. *Cognitive Science*, *38*(8), 1604–1633. https://doi.org/10.1111/cogs.12122

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology*, *52*(1), 31–45. https://doi.org/10.1037/dev0000065

Konovalov, A., & Krajbich, I. (2017). On the Strategic Use of Response Times. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3023640

Konovalov, A., & Krajbich, I. (2019). Revealed strength of preference: Inference from response times. *Judgment and Decision Making*, 14.

Köymen, B., Mammen, M., & Tomasello, M. (2016). Preschoolers use common ground in their justificatory reasoning with peers. *Developmental Psychology*, *52*(3), 423–429. https://doi.org/10.1037/dev0000089

Köymen, B., & Tomasello, M. (2018). Children's meta-talk in their collaborative decision making with peers. *Journal of Experimental Child Psychology*, *166*, 549–566. https://doi.org/10.1016/j.jecp.2017.09.018

Köymen, B., & Tomasello, M. (2020). The Early Ontogeny of Reason Giving. *Child Development Perspectives*, *14*(4), 215–220. https://doi.org/10.1111/cdep.12384

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*(1), 7455. https://doi.org/10.1038/ncomms8455

Laan, A., Madirolas, G., & de Polavieja, G. G. (2017). Rescuing Collective Wisdom when the Average Group Opinion Is Wrong. *Frontiers in Robotics and AI*, *4*. https://doi.org/10.3389/frobt.2017.00056

Laland, K. N. (2004). Social learning strategies. *Learning & Behavior*, 11. https://doi.org/doi.org/10.3758/BF03196002

Larson, J. R. (2010). *In search of synergy in small group performance*. Psychology Press.

Laughlin, P. R. (2011). *Group Problem Solving*. Princeton University Press.

Laughlin, P. R., Bonner, B. L., & Altermatt, T. W. (1998). Collective versus individual induction with single versus multiple hypotheses. *Journal of Personality and Social Psychology*, *75*(6), 1481–1489. https://doi.org/10.1037/0022-3514.75.6.1481

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on Letters-to-Numbers problems. *Organizational Behavior and Human Decision Processes*, 16. https://doi.org/10.1016/S0749-5978(02)00003-1

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and Social Combination Processes on Mathematical Intellective Tasks. *Journal of Experimental Social Psychology*, *22*, 177–189. https://doi.org/doi.org/10.1016/0022-1031(86)90022-3

Laughlin, P. R., & Futoran, G. C. (1985). Collective induction: Social combination and sequential transition. *Journal of Personality and Social Psychology*, *48*(3), 608–613. https://doi.org/10.1037/0022-3514.48.3.608

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology*, *90*(4), 644–651. https://doi.org/10.1037/0022-3514.90.4.644

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212. https://doi.org/10.1016/j.jecp.2014.03.001

Leonard, J. A., Bennett-Pierre, G., & Gweon, H. (2019). Who is better? Preschoolers infer relative competence based on efficiency of process and quality of outcome. *Proceedings of the Cognitive Science Society*, *7*.

Leonard, J. A., Garcia, A., & Schulz, L. E. (2019). How Adults' Actions, Outcomes, and Testimony Affect Preschoolers' Persistence. *Child Development*. https://doi.org/10.1111/cdev.13305

Leonard, J. A., Lee, Y., & Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290–1294. https://doi.org/10.1126/science.aan2317

Lerman, K., Yan, X., & Wu, X.-Z. (2016). The "Majority Illusion" in Social Networks. *PLOS ONE*, *11*(2), e0147617. https://doi.org/10.1371/journal.pone.0147617

Levine, M. E., & Plott, C. R. (1977). Agenda Influence and Its Implications. *Virginia Law Review*, *63*(4), 561. https://doi.org/10.2307/1072445

Liang, D. W., Moreland, R., & Argote, L. (1995). Group Versus Individual Training and Group Performance: The Mediating Role of Transactive Memory. *Personality and Social Psychology Bulletin*, *21*(4), 384–393. https://doi.org/10.1177/0146167295214009

Liberman, Z., & Shaw, A. (2020). Even his friend said he's bad: Children think personal alliances bias gossip. *Cognition*, *204*, 104376. https://doi.org/10.1016/j.cognition.2020.104376

Light, N., Fernbach, P. M., Rabb, N., Geana, M. V., & Sloman, S. A. (2022). Knowledge overconfidence is associated with anti-consensus views on controversial scientific issues. *Science Advances*, *8*(29), eabo0038. https://doi.org/10.1126/sciadv.abo0038

List, C., & Goodin, R. E. (2001). Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, *9*(3), 277–306. https://doi.org/10.1111/1467-9760.00128

List, C., & Pettit, P. (2006). Group Agency and Supervenience. *The Southern Journal of Philosophy*, *44*(S1), 85–105. https://doi.org/10.1111/j.2041-6962.2006.tb00032.x

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041. https://doi.org/10.1126/science.aag2132

Lombrozo, T. (2016). Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences*, *20*(10), 748–759. https://doi.org/10.1016/j.tics.2016.08.001

Looser, C. E., & Wheatley, T. (2010). The Tipping Point of Animacy: How, When, and Where We Perceive Life in a Face. *Psychological Science*, *21*(12), 1854–1862. https://doi.org/10.1177/0956797610388044

Magid, R. W., Yan, P., Siegel, M. H., Tenenbaum, J. B., & Schulz, L. E. (2018). Changing minds: Children's inferences about third party belief revision. *Developmental Science*, *21*(2), e12553. https://doi.org/10.1111/desc.12553

Mahr, J. B., & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X17000012

Mahr, J. B., & Csibra, G. (2022). A Short History of Theories of Intuitive Theories. In J. Gervain, G. Csibra, & K. Kovács (Eds.), *A Life in Cognition* (Vol. 11, pp. 219–232). Springer International Publishing. https://doi.org/10.1007/978-3-030-66175-5_16

Mannes, A. E. (2009). Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science*, *55*(8), 1267–1279. https://doi.org/10.1287/mnsc.1090.1031

Marks, G., & Miller, N. (1987). Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review. *Psychological Bulletin*, *102*(1), 19.

Mason, W. A., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, *137*(3), 422–433. https://doi.org/10.1037/a0012798

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, *109*(3), 764–769. https://doi.org/10.1073/pnas.1110069108

Massoni, S., & Roux, N. (2017). Optimal group decision: A matter of confidence calibration. *Journal of Mathematical Psychology*, *79*, 121–130. https://doi.org/10.1016/j.jmp.2017.04.001

Mastroianni, A. M., Gilbert, D. T., Cooney, G., & Wilson, T. D. (2021). Do conversations end when people want them to? *Proceedings of the National Academy of Sciences*, *118*(10), e2011809118. https://doi.org/10.1073/pnas.2011809118

Meltzoff, A. N., & Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology*, *44*(5), 1257–1265. https://doi.org/10.1037/a0012888

Mercier, H. (2016). The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

Mercier, H., & Claidière, N. (2022). Does discussion make crowds any wiser? *Cognition*, *222*, 104912. https://doi.org/10.1016/j.cognition.2021.104912

Mercier, H., Dockendorff, M., Majima, Y., Hacquin, A.-S., & Schwartzberg, M. (2020). Intuitions about the epistemic virtues of majority voting. *Thinking & Reasoning*, 1–19. https://doi.org/10.1080/13546783.2020.1857306

Mercier, H., & Miton, H. (2019). Utilizing simple cues to informational dependency. *Evolution and Human Behavior*, *40*(3), 301–314. https://doi.org/10.1016/j.evolhumbehav.2019.01.001

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Mercier, H., & Sperber, D. (2019). Précis of The Enigma of Reason. *Teorema*, *38*, 8.

Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning*, *21*(3), 341–355. https://doi.org/10.1080/13546783.2014.981582

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, *135*(5), 749–773. https://doi.org/10.1037/a0016854

Miller, N., Garnier, S., Hartnett, A. T., & Couzin, I. D. (2013). Both information and social cohesion determine collective decisions in animal groups. *Proceedings of the National Academy of Sciences*, *110*(13), 5263–5268. https://doi.org/10.1073/pnas.1217513110

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology*, *49*(3), 404–418. https://doi.org/10.1037/a0029500

Mills, C. M., Al-Jabari, R. M., & Archacki, M. A. (2012). Why do People Disagree? Explaining and Endorsing the Possibility of Partiality in Judgments. *Journal of Cognition and Development*, *13*(1), 111–136. https://doi.org/10.1080/15248372.2010.547236

Mills, C. M., & Grant, M. G. (2009). Biased decision-making: Developing an understanding of how positive and negative relationships may skew judgments. *Developmental Science*, *12*(5), 784–797. https://doi.org/10.1111/j.1467-7687.2009.00836.x

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87*(1), 1–32. https://doi.org/10.1016/j.jecp.2003.09.003

Mills, C. M., & Keil, F. C. (2005). The Development of Cynicism. *Psychological Science*, *16*(5), 385–390. https://doi.org/10.1111/j.0956-7976.2005.01545.x

Mills, C. M., & Keil, F. C. (2008). Children's developing notions of (im)partiality. *Cognition*, *107*(2), 528–551. https://doi.org/10.1016/j.cognition.2007.11.003

Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 639–648. https://doi.org/10.1098/rstb.2006.2000

Morgan, T. J. H., & Laland, K. N. (2012). The Biological Bases of Conformity. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00087

Morgan, T. J. H., Laland, K. N., & Harris, P. L. (2015). The development of adaptive conformity in young children: Effects of uncertainty and consensus. *Developmental Science*, *18*(4), 511–524. https://doi.org/10.1111/desc.12231

Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1729), 653–662. https://doi.org/10.1098/rspb.2011.1172

Moshman, D., & Geil, M. (1998). Collaborative Reasoning: Evidence for Collective Rationality. *Thinking & Reasoning*, *4*(3), 231–248. https://doi.org/10.1080/135467898394148

Muthukrishna, M., Morgan, T. J. H., & Henrich, J. (2016). The when and who of social learning and conformist transmission. *Evolution and Human Behavior*, *37*(1), 10–20. https://doi.org/10.1016/j.evolhumbehav.2015.05.004

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126–132. https://doi.org/10.1038/s41562-017-0273-4

Nicholls, J. G. (1978). The Development of the Concepts of Effort and Ability, Perception of Academic Attainment, and the Understanding That Difficult Tasks Require More Ability. *Child Development*, *49*(3), 16.

Nokes-Malach, T. J., Meade, M. L., & Morrow, D. G. (2012). The effect of expertise on collaborative problem solving. *Thinking & Reasoning*, *18*(1), 32–58. https://doi.org/10.1080/13546783.2011.642206

Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When Is It Better to Learn Together? Insights from Research on Collaborative Learning. *Educational Psychology Review*, *27*(4), 645–656. https://doi.org/10.1007/s10648-015-9312-8

Novaes, C. D. (2020). *The Dialogical Roots of Deduction: Historical, Cognitive, and Philosophical Perspectives on Reasoning*. Cambridge University Press.

Noyes, A., Gerdin, E., Rhodes, M., & Dunham, Y. (2023). *A developmental investigation of group concepts in the context of social hierarchy: Can the powerful impose group membership?* (Preprint PsyArXiv). PsyArXiv. https://doi.org/10.31234/osf.io/gjtxv

Oktar, K., & Lombrozo, T. (2022). Mechanisms of Belief Persistence in the Face of Societal Disagreement. *Proceedings of the Cognitive Science Society*, 8.

O'Madagain, C., & Tomasello, M. (2019). Joint attention to mental content and the social origin of reasoning. *Synthese*. https://doi.org/10.1007/s11229-019-02327-1

Orena, A. J., & White, K. S. (2015). I Forget What That's Called! Children's Online Processing of Disfluencies Depends on Speaker Knowledge. *Child Development*, *86*(6), 1701–1709. https://doi.org/10.1111/cdev.12421

Papert, S. (1966). *The Summer Vision Project* (Project Mac AI Memo 100). IT. https://dspace.mit.edu/handle/1721.1/6125

Perret, P., & Dauvier, B. (2018). Children's Allocation of Study Time during the Solution of Raven's Progressive Matrices. *Journal of Intelligence*, *6*(1), 9. https://doi.org/10.3390/jintelligence6010009

Perret-Clermont, A.-N., Carugati, F., & Oates, J. (2004). A Socio-cognitive perspective on learning and cognitive development. In *Cognitive and Language Development in Children*. Blackwell Publishing Ltd.

Pham, T., & Buchsbaum, D. (2020). Children's use of majority information is influenced by pragmatic inferences and task domain. *Developmental Psychology*, *56*(2), 312–323. https://doi.org/10.1037/dev0000857

Pietraszewski, D. (2022). Toward a computational theory of social groups: A finite set of cognitive primitives for representing any and all social groups in the context of conflict. *Behavioral and Brain Sciences*, *45*, e97. https://doi.org/10.1017/S0140525X21000583

Pietraszewski, D., Curry, O. S., Petersen, M. B., Cosmides, L., & Tooby, J. (2015). Constituents of political cognition: Race, party politics, and the alliance detection system. *Cognition*, *140*, 24–39. https://doi.org/10.1016/j.cognition.2015.03.007

Pratt, S. C., & Sumpter, D. J. T. (2006). A tunable algorithm for collective decision-making. *Proceedings of the National Academy of Sciences*, *103*(43), 15906–15910. https://doi.org/10.1073/pnas.0604801103

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535. https://doi.org/10.1038/nature21054

Pun, A., Birch, S. A. J., & Baron, A. S. (2016). Infants use relative numerical group size to infer social dominance. *Proceedings of the National Academy of Sciences*, *113*(9), 6. https://doi.org/10.1073/pnas.1514879113

Putnam, H. (1975). The Meaning of "Meaning." In *Minnesota studies in the philosophy of science: Language, Mind, and Knowledge* (Vol. 7, pp. 131–193). University of Minnesota Press.

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*(10), 420–428. https://doi.org/10.1016/j.tics.2009.08.002

Rabb, N., Fernbach, P. M., & Sloman, S. A. (2019). Individual Representation in a Community of Knowledge. *Trends in Cognitive Sciences*, *23*(10), 891–902. https://doi.org/10.1016/j.tics.2019.07.011

Rand, D. G. (2016). Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. *Psychological Science*, *27*(9), 1192–1206. https://doi.org/10.1177/0956797616654455

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Richardson, E., Sheskin, M., & Keil, F. C. (2021). An Illusion of Self-Sufficiency for Learning About Artifacts in Scaffolded Learners, But Not Observers. *Child Development*, 16. https://doi.org/10.1111/cdev.13506

Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication*, *48*(9), 1079–1093. https://doi.org/10.1016/j.specom.2006.02.001

Ross, C. T., Winterhalder, B., & McElreath, R. (2018). Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police. *Palgrave Communications*, *4*(1), 61. https://doi.org/10.1057/s41599-018-0110-z

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Salthouse, T. A. (1996). The Processing-Speed Theory of Adult Age Differences in Cognition. *Psychological Review*, *103*(3), 26. https://doi.org/doi.org/10.1037/0033-295X.103.3.403

Sasaki, T., Stott, B., & Pratt, S. C. (2019). Rational time investment during collective decision making in *Temnothorax* ants. *Biology Letters*, *15*(10), 20190542. https://doi.org/10.1098/rsbl.2019.0542

Schmidt, M. F. H., Rakoczy, H., Mietzsch, T., & Tomasello, M. (2016). Young Children Understand the Role of Agreement in Establishing Arbitrary Norms-But Unanimity Is Key. *Child Development*, *87*(2), 612–626. https://doi.org/10.1111/cdev.12510

Scott, R. M., & Baillargeon, R. (2017). Early False-Belief Understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Sears, D. A., & Reagin, J. M. (2013). Individual versus collaborative problem solving: Divergent outcomes depending on task complexity. *Instructional Science*, *41*(6), 1153–1172. https://doi.org/10.1007/s11251-013-9271-8

Sheskin, M., & Keil, F. (2018). TheChildLab.com A Video Chat Platform for Developmental Research. *PsyArxiv*. https://doi.org/10.31234/osf.io/rn7w5

Siegel, M. H., Magid, R., Tenenbaum, J. B., & Schulz, L. E. (2014). Black boxes: Hypothesis testing via indirect perceptual evidence. *Proceedings of the Cognitive Science Society*, 7.

Singer, M., & Tiede, H. L. (2008). Feeling of knowing and duration of unsuccessful memory search. *Memory & Cognition*, *36*(3), 588–597. https://doi.org/10.3758/MC.36.3.588

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why Peer Discussion Improves Student Performance on In-Class Concept Questions. *Science*, *323*, 4. https://doi.org/10.1126/science.1165919

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*(1), B1–B11. https://doi.org/10.1016/j.cognition.2004.07.004

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-Detection Analysis of Group Decision Making. *Psychological Review*, *108*(1), 21. https://doi.org/10.1037/0033-295X.108.1.183

Stasser, G., & Abele, S. (2020). Collective Choice, Collaboration, and Communication. *Annual Review of Psychology*, *71*(1), 589–612. https://doi.org/10.1146/annurev-psych-010418-103211

Stasser, G., & Stewart, D. (1992). Discovery of Hidden Profiles by Decision-Making Groups: Solving a Problem Versus Making a Judgment. *Journal of Personality and Social Psychology*, *63*(3), 426–434. https://doi.org/10.1037/0022-3514.63.3.426

Stasser, G., & Titus, W. (2003). Hidden Profiles: A Brief History. *Psychological Inquiry*, *14*(3–4), 304–313. https://doi.org/10.1080/1047840X.2003.9682897

Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, *9*(4), 465–481. https://doi.org/10.1007/s11097-010-9174-y

Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G., & Plotkin, J. B. (2019). Information gerrymandering and undemocratic decisions. *Nature*, *573*(7772), 117–121. https://doi.org/10.1038/s41586-019-1507-6

Steyvers, M., Miller, B., Hemmer, P., & Lee, M. D. (2009). The Wisdom of Crowds in the Recollection of Order Information. *Advances in Neural Information Processing Systems*, 9.

Stipek, D., & Iver, D. M. (1989). Developmental Change in Children's Assessment of Intellectual Competence. *Child Development*, *60*(3), 19.

Sulik, J., Bahrami, B., & Deroy, O. (2020). Social influence and informational independence. *Proceedings of the Cognitive Science Society*, 7. https://cognitivesciencesociety.org/cogsci20/papers/0704/0704

Sumpter, D. J. T., & Pratt, S. C. (2009). Quorum responses and consensus decision making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 743–753. https://doi.org/10.1098/rstb.2008.0204

Teasley, S. D. (1995). The Role of Talk in Children's Peer Collaborations. *Developmental Psychology*, *31*(2), 14. https://doi.org/10.1037/0012-1649.31.2.207

Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., & Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, *119*(4), e2116915119. https://doi.org/10.1073/pnas.2116915119

Theiner, G. (2013). Transactive Memory Systems: A Mechanistic Analysis of Emergent Group Memory. *Review of Philosophy and Psychology*, *4*(1), 65–89. https://doi.org/10.1007/s13164-012-0128-x

Theiner, G., Allen, C., & Goldstone, R. L. (2010). Recognizing group cognition. *Cognitive Systems Research*, *11*(4), 378–395. https://doi.org/10.1016/j.cogsys.2010.07.002

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis. *Current Anthropology*, *53*(6), 673–692. https://doi.org/10.1086/668207

Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, *3*(2), 183–193. https://doi.org/10.1038/s41562-018-0518-x

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

van Leeuwen, E. J. C., Cohen, E., Collier-Baker, E., Rapold, C. J., Schäfer, M., Schütte, S., & Haun, D. B. M. (2018). The development of human social learning across seven societies. *Nature Communications*, *9*(1), 2076. https://doi.org/10.1038/s41467-018-04468-2

von Hippel, W., Ronay, R., Baker, E., Kjelsaas, K., & Murphy, S. C. (2016). Quick Thinkers Are Smooth Talkers: Mental Speed Facilitates Charisma. *Psychological Science*, *27*(1), 119–122. https://doi.org/10.1177/0956797615616255

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357. https://doi.org/10.1016/j.cognition.2014.07.008

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining Constrains Causal Learning in Childhood. *Child Development*, *88*(1), 229–246. https://doi.org/10.1111/cdev.12590

Ward, A. J. W., Sumpter, D. J. T., Couzin, I. D., Hart, P. J. B., & Krause, J. (2008). Quorum decision-making facilitates information transfer in fish shoals. *Proceedings of the National Academy of Sciences*, *105*(19), 6948–6953. https://doi.org/10.1073/pnas.0710344105

Wearden, J. (2016). *The psychology of time perception*. Palgrave Macmillan.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior* (pp. 185–208). Springer. http://link.springer.com/chapter/10.1007/978-1-4612-4634-3_9

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Whalen, A., Griffiths, T. L., & Buchsbaum, D. (2018). Sensitivity to Shared Information in Social Learning. *Cognitive Science*, *42*(1), 168–187. https://doi.org/10.1111/cogs.12485

Whiten, A. (2017). Social Learning and Culture in Child and Chimpanzee. *Annual Review of Psychology*, *68*(1), 129–154. https://doi.org/10.1146/annurev-psych-010416-044108

Williams, J. J., & Lombrozo, T. (2010). The Role of Explanation in Discovery and Generalization: Evidence From Category Learning. *Cognitive Science*, *34*(5), 776–806. https://doi.org/10.1111/j.1551-6709.2010.01113.x

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*(4), 1006–1014. https://doi.org/10.1037/a0030996

Wohltjen, S., & Wheatley, T. (2021). Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences*, *118*(37), e2106645118. https://doi.org/10.1073/pnas.2106645118

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147

Xie, B., & Hayes, B. (2022). Sensitivity to Evidential Dependencies in Judgments Under Uncertainty. *Cognitive Science*, *46*(5). https://doi.org/10.1111/cogs.13144

Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The Illusion of Consensus: A Failure to Distinguish Between True and False Consensus. *Psychological Science*, *30*(8), 1195–1204. https://doi.org/10.1177/0956797619856844