Herding cats: children and adults infer collective decision speed from team size and diversity, but disagree about whether consensus strength matters more than team size

Emory Richardson (rchrdsn.emory@gmail.com), **Hannah Hok** (hok.uchicago@gmail.com)
Alex Shaw (ashaw1@uchicago.edu), & **Frank Keil** (frank.keil@yale.edu)

Abstract

Collaboration can make collective judgments more accurate than individual judgments, but it also comes with costs in time, effort, and social cohesion. Here we focus on time costs. 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 power and consensus in group dynamics.

Keywords: consensus, decision speed, collective behavior, social cognition, conceptual development, intuitive theories

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 can 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 (Almaatouq et al., 2021; Goldstone et al., 2023; Tomasello et al., 2012). By adulthood, it seems commonsensical that collaborators need to weigh the *costs* of deliberation as well as 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. Our question here is how people estimate the time costs of debate.

The remainder of the introduction is structured as follows. The first section is a theoretical justification: why study people's intuitions about group decision speed, and what makes children's inferences particularly revealing? Importantly, these intuitions aren't simply illusions: they're endogenous constraints on collective decisions, confirmed in simulations and empirical studies. The second section lays out our predictions about adults' inferences. The third section explains why children's inferences may differ.

Why study intuitions about group decision speed, and why in children?

In short, because even though the social dynamics that drive collective decision-making are clearly complex, reasoning about how they contribute to decision speed doesn't seem to require much effort — and seemingly effortless inferences about complex phenomena are a hallmark of intuitive theories. Intuitive theories are a sparse framework of a few salient cues and some beliefs about their causal connections that are thought to guide conceptual development and shape adult reasoning about the natural world (Keil, 2011; Ullman & Tenenbaum, 2020; Mahr & Csibra, 2022). Importantly, intuitive theories don't need to be particularly accurate or precise. They simply need to allow people to navigate a conceptual domain in everyday life, and be flexible enough to accommodate cognitive development and conceptual change.

For instance, it seems commonsensical that large groups will take longer to make decisions than small groups, or that groups in which a strong initial consensus can pressure dissenters to concede will make decisions more quickly than groups evaluating multiple competing perspectives with no initial consensus at all. Why? We suggest that people's inferences about group decision speed 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. The component intuitions we focus on here are: (1) expressing an opinion takes time, (2) debating differences takes even more, and (3) while 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 1). Talk may take more time when there are more opinions to express or debate, but any teammate can *stop* talking whenever they want; someone who is willing to simply abide by any group decision doesn't have to take up airtime. However, one person's unilateral withdrawal is only guaranteed to *save* time in Panel 1, where the debate will end as soon as either teammate acquiesces. By contrast, out of the five teammates in Panel 2, only the singleton can end the debate unilaterally by acquiescing: after all, even if one of the other four withdrew, 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 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 more coordination required to make a decision, the longer it will take; and decisions require more coordination in teams with more people, more factions, and a more balanced distribution of power.

Critically, these intuitions aren't simply illusions. Agent-based simulations demonstrate that increasing a group's size or diversity of preferences lead to slower decisions, while lower decision thresholds (e.g., plurality or majority instead of supermajority or unanimity) can speed up decisions — in other words, mutually acknowledged deference to the proportionally largest faction can short-circuit endless dissent (Albrecht, Anderson, & Vroman, 2010; Chan, Lizzeri,

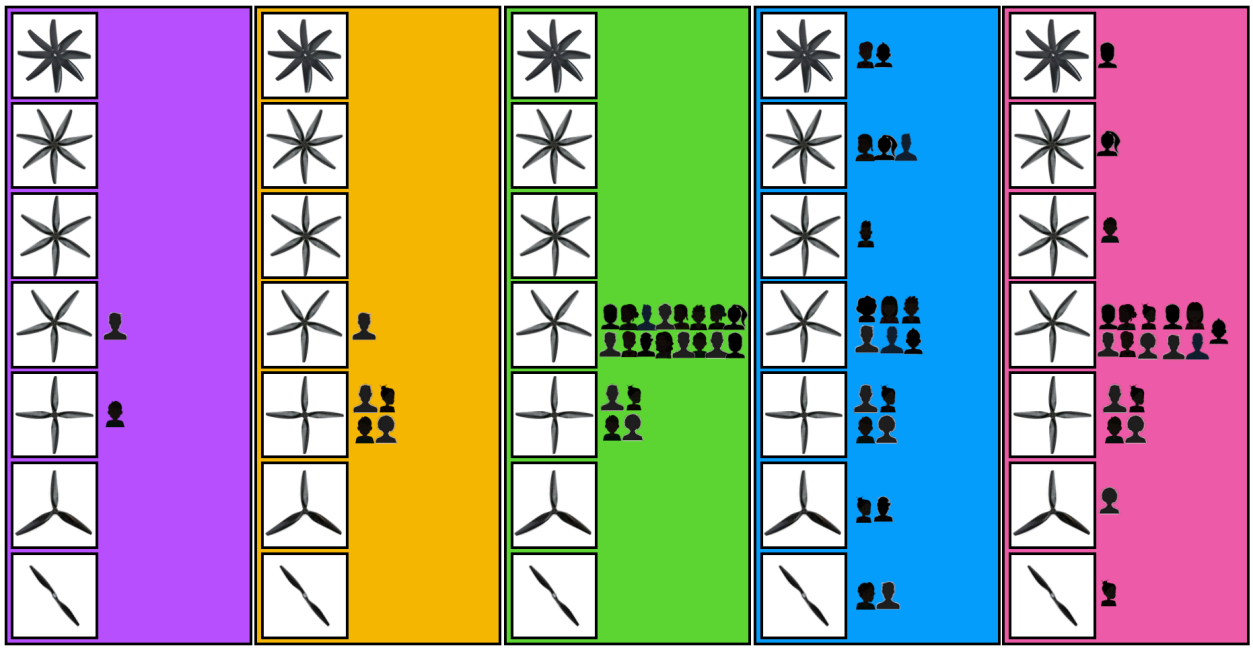


Figure 1. Five different robotics teams divided into two or more factions. Since they disagree, they need to talk together to make a decision about which propeller to use. Panels 2-3 have the same proportional distributions; teams in Panels 3-5 are the same size, but the faction endorsing the 4-blade varies in factional power across panels while representing the same proportion of teammates.

Suen, & Yariv, 2018; Stein, Frey, & Flache, 2024). Empirical studies tell the same story about collective decisions in other species, and suggest similar dynamics in both children and adults (Conradt & List, 2009; Kameda, Toyokawa, & Tindale, 2022; Kearns, 2012; Kearns, Suri, & Monfort, 2006; Brocas & Carrillo, 2024). Our point here is simply that if humans are equipped with intuitive theories that 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 that most strongly shape our intuitive theories may change from early development to adulthood.

Our experiments focus on three features of groups: the number of people ("size" for short), the number factions ("diversity" of opinions), and the relative proportions of the factions ("consensus strength"). We focus our experiments on ages 6 to 9. As we'll review in Section 3 below, the developmental literature provides indirect evidence that intuitions about how size and diversity constrain decision speeds may be in place by age 6; but, it also suggests that reasoning about the role of consensus in group judgments may continue to undergo conceptual change until age 9 or even later.

Our predictions are as follows. Following the basic intuitions that coordination takes more time when there are more people or factions to coordinate between, we predict that both adults and children of all ages will expect slower decisions when groups are larger or more diverse, and that the diversity will matter more than size when the two are contrasted. And following the intuition that stronger consensus puts more pressure on dissenters to concede, we predict that adults will infer quicker decisions from teams in which one faction is especially strong than from teams divided into more equally-sized factions. But, we predict that children may not

share the adult intuition; instead, we predict that in trials that contrast consensus strength with team size, children will expect slower decisions from larger teams, even if the consensus is weaker than on the smaller team. We focus the next section on adults' reasoning, to make each of these predictions (for size, diversity, and consensus strength) more clear. In Section 3, we explain when and why children's reasoning may differ.

Adult's intuitions

We predict that adults will expect slower decisions from teams with more people or factions, and expect the number of factions to matter more than the number of people — following straightforwardly from the basic intuitions that talk takes time (more people means more talk) and debate takes more (more factions means more debate). But we also predict that they'll 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). When team members are mutually aware that the group is moving towards consensus, dissenters may feel growing pressure to conform, even if they disagree — and making the strongest faction increasingly difficult to fracture. But in teams with multiple factions, the power one faction holds over another may depend not just on its size, but also on its relationship with other factions. In other words, group dynamics may often depend more on "party discipline" (i.e., how strictly individuals subordinate their idiosyncrasies to the interests of their own factions) and conformist tendencies (i.e., deference to consensus) of two minority factions than the proportional or numerical size of the largest faction. To the extent that group decision-making is constrained by these kinds of consensus-based power dynamics as well as the total number of people and factions, we would expect all three to be reflected in adults' intuitive theories. However, adults' intuitive theories are also shaped by conceptual development in early childhood.

Children's intuitions

We predict that young children, like adults, will expect slower decisions from teams with more people or factions. Why? Because talk takes time, and by age six children have at least two ways to infer *how much talk* goes into resolving disagreements in large and factious groups. First: reasoning about the relationship between time, effort, and task difficulty emerges early. Even four year olds expect more difficult *physical* tasks to take longer to complete (Leonard, Bennet-Pierre, & Gweon, 2019). But by age six, children are able to make similar inferences about more abstract tasks: they infer that more complex *reasoning* problems will take longer to solve, even when no physical cues are present (Richardson & Keil, 2022). We take resolving disagreement to be a complex reasoning task. Six year olds may infer that having more factions or people on a team makes coordination more complex, and therefore slower. Second: by age six, children may also be able to infer how much talk goes into resolving disagreements by drawing on their own

experience of collaborative reasoning. Even preschoolers explicitly dispute statements they believe to be false, and how much of their reasoning they verbalize depends on what they expect their collaborators to know already (Köymen, Mammen, & Tomasello, 2016). But as Tomasello (2021), a child's six or seventh birthday marks an inflection point — what many cultures have traditionally considered the "age of reason" — after which children begin to use strategies for engaging with peers more deeply and efficiently, such as engaging in meta-talk comparing their relative confidence or their informants' reliability as sources (Köymen & Tomasello, 2018). Along with believing that larger groups make coordination more complex and time-consuming, children's own experience of being increasingly efficient collaborators could make them 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; van Leeuwen et al., 2018; Pham & Buchsbaum, 2020). Children don't simply defer more — they become more discerning about whether or not to defer at all. 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 less skilled than adults in adjudicating conflicting explanations, and often strikingly overconfident in their own knowledge (Kloo, Rohwer, & Perner, 2017; Mills & Keil, 2004). For instance, while preschoolers do evaluate each others' reasoning, they only begin to engage in meta-talk—such as comparing confidence levels or informant reliability—upon reaching the "age of reason" at age 6 or 7 (Köymen & Tomasello, 2018; Tomasello, 2020). 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 weigh team size more heavily than the distribution of factional power. If so, they may infer that a *large* team with *strong* initial consensus will still take longer to make decisions than a *small* team with *little or no* initial consensus.

To be clear, the claim is not that children fail to notice differences in consensus strength at all. 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 may expect an agent with one physically large ally to make way on a narrow bridge for an agent with two smaller allies whose cumulative size is equal to the larger (Pun, Birch, & Baron, 2016; but see Yousif & Keil, 2021). 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 usually vested in the few rather than the many (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. But in Heck et al. (2021), children's inferences didn't scale with size for the strength-in-numbers task (even though, like adults, they were more likely to attribute *authority* to proportionally smaller groups). And Pun et al.'s (2016) studies weren't designed to test whether power scaled with proportional differences (infants only saw groups of 3 and 2). 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.

In two pre-registered experiments[1], 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.

Experiment 1

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. Thus, in a second task following the experiment (*Build*), we ask which of two teams (20v10) would take longer to physically *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

[1] Link to pre-registrations, materials, power analyses, data:https://osf.io/9xtyu/?view_only=037914869b2c43b2bfeff1a3e4134bb7

once consensus is reached. The *Contrast* and *Build* 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 than a smaller team. Pre-registrations and other materials, including a power analysis, can be found in the first author's OSF repository[2]. As explained in our pre-registration, we conducted a power analysis by simulation in order to compare robustness to different effect sizes and a potential effect of counterbalance observed in the pilot data. However, no counterbalance effects were observed in the full experiment.

Participants. Based on a power analysis simulation, 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. Though we no longer have access to participant-specific demographics (see SI materials), the participant database at the start of Experiment 1 included ZIP codes from all 50 U.S. states with a median yearly income of $54,172, and reported racial demographics (including multi-racial) of 72% White (60% reporting White + no other categories), 10% Hispanic or Latino, 7% Black or African-American, 6% Asian, 2% American Indian or Alaska Native, 2% Other, 1% Native Hawaiian or Other Pacific Islander.

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?*" essentially presenting a 7-point Likert scale in two-stages (for similar uses of Likert scales with children, see Bass et al., 2022; Mills & Keil, 2004; Ahl, Amir, & McAuliffe, 2024; DeJesus, Venkatesh, & Kinzler, 2021; Lapidow, Killeen, & Walker, 2021). Adults responded using the same 7-point scale directly. 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

[2] Link to pre-registration, materials, power analysis, data: https://osf.io/9xtyu/?view_only=037914869b2c43b2bfeff1a3e4134bb7

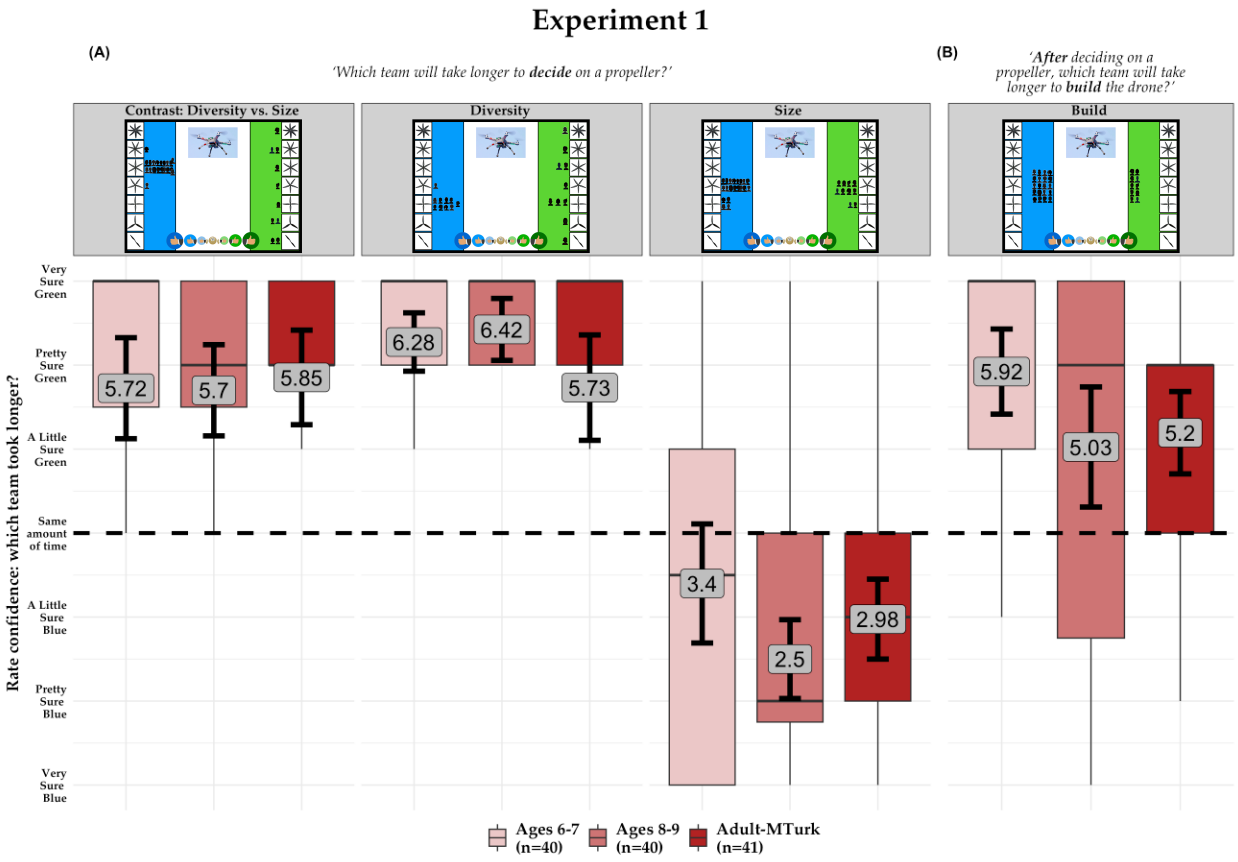


Figure 2. Box plots showing results from Exp 1. Box plot shading indicates age group; grey labels display means, error bars are 95% CIs. Facets display last slide of given trial. **(A) Decision time**: each participant rated each trial, in counter-balanced order. **(B) Build time**: each participant rated build time after completing all three decision trials.

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.

Results and Discussion. 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 to our pre-registered analysis plan. This makes the intercept equivalent to a 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*. Adults and the child sample as a whole expected reaching consensus decisions about how to build a drone to take longer in larger teams than smaller teams, but physically building one after deciding to take less time (*Size* trial: $\beta_{Intercept}$ = -1.05, SE = .21, p<.0001, [95CI: -1.46, -0.64]; *Build* trial: $\beta_{Intercept}$ = 1.48, SE = .22, p<.0001, [95CI: 1.04, 1.91]). However, they also inferred that reaching consensus would take longer in teams divided into more factions, regardless of whether the more factious team was the same size (*Diversity* trial: $\beta_{Intercept}$ = 2.35, SE = .13, p<.001, [95CI: 2.10, 2.60]) or smaller than the less

factious team (*Contrast* trial: $\beta_{Intercept}$ = 1.71, SE = .20, p<.0001, [95CI: 1.31, 2.11]). No age effects were observed in the *Diversity* or *Contrast* trials; however, age was 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. 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]); all other t.tests supported our primary analysis.

What do these results tell us about participants' reasoning process? First, participants weren't simply mapping a "more time" response to the team with more people or more factions; if they were, they wouldn't have expected team size to have opposite effects on cognitive decision speed (*Size* trial) and physical build speed (*Build* trial). A more-is-more heuristic also doesn't explain why participants would expect decision speed to depend more on the number of factions than the number of people (*Contrast* trial). Second, the difference between the physical task in the Build trial and the decision task in the other three trials suggests that participants' inferences specifically reflected their beliefs about how teams make consensus *decisions*. But Experiment 1 alone doesn't tell us what those beliefs are. For instance, one might simply assume *an* outcome (either majority rule, or whichever propeller seemed best to the participant themselves), and 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). Thus, in Experiment 2, we ask participants to infer which team would take longer given that *both* teams chose the *same* propeller. This allows us to control for the numerical and proportional size of the winning and losing factions, whether any faction constituted a majority at the outset, and the total number of people and factions. We also chose to tell participants that the propeller chosen by the winning factions was in fact optimal. While it's possible that our results would differ if we said the winning faction was *in*accurate, majority judgment is often both a default decisions rule and a cue to accuracy; thus, since we're already manipulating the strength of the winning faction, we decided it would be simpler to not contradict common assumptions about majority accuracy.

Experiment 2

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 we also predict that while adults will expect consensus strength to matter more than size, children will infer just the opposite, as explained in the introduction. For instance, while 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) inferences about size may not differ from chance even though they do differ for teams with more factions, our preregistration treats older children as the primary developmental contrast for the trials in which size and factional power are contrasted. More specifically, we pre-register separate regressions for younger and older children on each trial, and our hypotheses focus on the older children's responses. Younger children may show the same pattern as older children; but if they do not differ from chance, further work would be needed to understand why. As with Experiment 1, materials and pre-registrations for Exp 2 are available at the first author's OSF repository[3]. We again used a simulation for our power analysis in order to be able to visualize and better understand various plausible effect sizes for each age group.

Participants. Based on a power analysis simulating various effect sizes that seemed plausible based on pilot data, 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. Though we no longer have access to participant-specific demographics (see SI materials), the participant database at the start of Experiment 1 included ZIP codes from all 50 U.S. states with a median yearly income of $54,172, and reported racial demographics (including multi-racial) of 72% White (60% reporting White + no other categories), 10% Hispanic or Latino, 7% Black or African-American, 6% Asian, 2% American Indian or Alaska Native, 2% Other, 1% Native Hawaiian or Other Pacific Islander.

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,

[3] Link to pre-registration, materials, power analyses, data: https://osf.io/9xtyu/?view_only=037914869b2c43b2bfeff1a3e4134bb7

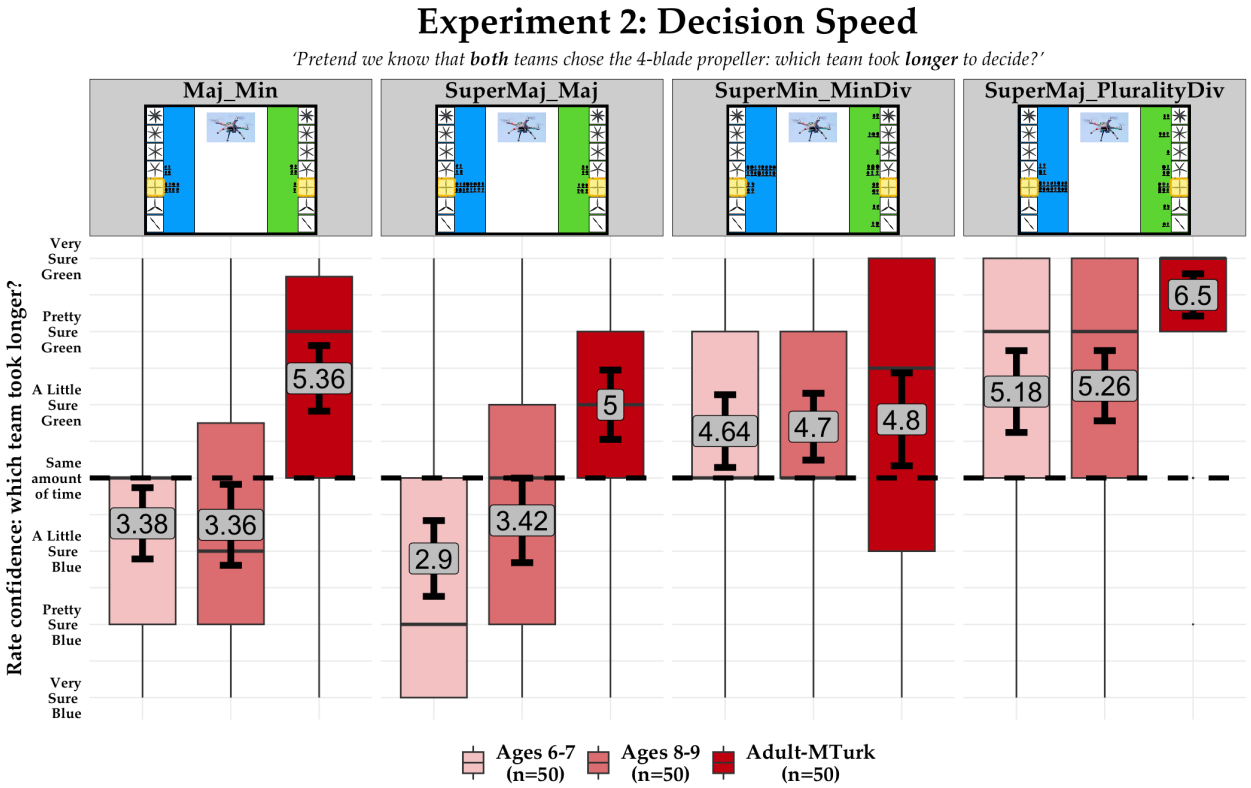


Figure 3. Box plots showing results from decision speed task in Experiment 2. Participants were told that the propeller that was "actually the best" was the 4-blade propeller (highlighted in yellow), and asked to pretend that both teams chose the best propeller after talking together. Facets display critical slide from the procedure for each trial; each participant rated each trial, in counter-balanced order. Box plot shading indicates age group; grey labels display means, error bars are 95% CIs.

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 or not the team would decide to use the 4-blade propeller after talking.

Results and Discussion. Experiment 2 provides direct evidence against a number of heuristics simpler than the kind of reasoning about disagreement we've proposed. 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".

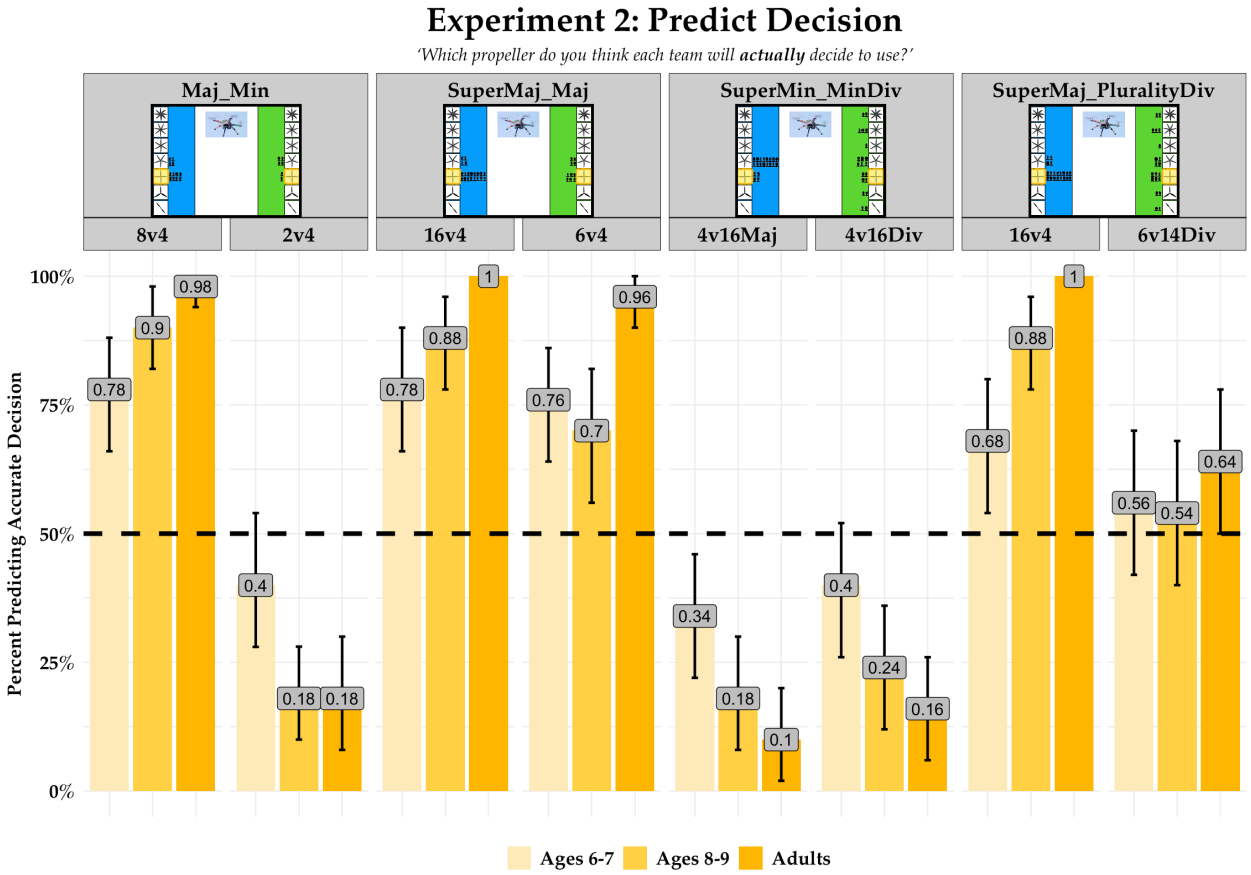


Figure 4. Bar plot showing choice predictions in Experiment 2. After inferring which team in each trial would have taken longer if both teams had ultimately chosen the 4-blade propeller described participants were told was "actually best" (highlighted in yellow), participants were asked which propeller they thought each team would *actually* choose, as a forced-choice between the 4-blade propeller and any other propeller. Grey labels show the percentage predicting an accurate decision; error bars are 95% CIs; bar shading displays age groups. Facets are nested to show the predictions for each team in each trial.

On the *SuperMaj_PluralityDiv* and *SuperMin_MinDiv* trials, each team had 20 teammates. And 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-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, η_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} = 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 (e.g., the team with no initial majority) would need any time at all to make a decision.

Finally, when asked to predict each team's final decision in the second task, participants tended to expect the proportionally largest faction to prevail regardless of whether or not the propeller that faction had endorsed was the best option (Figure 4.4; SI materials Table 1). 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 it.

General Discussion

Consensus doesn't come from the group simply figuring out what's best. It's often negotiated, expedient, and costly to achieve. One cost is time. But coordinating consensus decisions in groups means that individuals yield unilateral control over the time they spend on a decision as well as the decision itself. Instead, a decision's speed and accuracy both depend on social dynamics. So managing speed-accuracy tradeoffs in groups means that collaborators need to know how to pick their battles. Taken together, our experiments suggest that some of the intuitions that help people decide which battles are worth the time emerge in early childhood — but they may also change as a result of conceptual development.

Like adults, children as young as six expected slower decisions from teams with more people or more factions. This wasn't because participants thought larger teams do *everything* more slowly: all ages said that *building* drones would be faster in teams with *more* people once the team had decided on a design. But children didn't appear to share adults' intuition that dividing a team into proportionally unequal factions would speed up consensus-congruent decisions (and slow down consensus-incongruent decisions): in the Experiment 2 trials that

contrasted two-faction teams but controlled the size of the "losing" factions (8v4-and-2v4 or 16v4-and-6-v4), children predicted slower decisions from the larger team despite it having a stronger initial consensus.

We doubt children's size-over-strength inferences reflect a failure of proportional reasoning: even preschoolers can easily distinguish the vote ratios we used (2:1, 3:2, 4:1) in the two faction trials (Halberda & Feigenson, 2008). And children did *predict* majority-rule, suggesting that they didn't have trouble recognizing how votes were initially proportioned — they simply didn't expect consensus strength to influence decision speeds more than team size. Children's inferences in Experiment 2 also provide evidence against a variety of simple heuristic strategies (e.g., "more is more", assuming majority rule, number of winners, number of losers, etc). So why don't children take consensus strength into account?

Our suggestion, put briefly, has been that one reason *adults* may expect decision speed to depend less on group size than consensus strength is because the intuitive theories that allow people to manage group dynamics include a mutual expectation of stronger deference to stronger consensus. Stronger deference to stronger consensus can short-circuit endless dissent in groups of any size — but it's a mechanism that that only works at all to the extent it works on *everyone*. After all, time spent is time spent, regardless of whether it's spent wrangling a single stubborn dissenter or working through a multi-faction negotiation. And while even preschoolers do defer to majorities, existing evidence suggests that strength-based deference to consensus is only beginning to emerge around ages 6-7 (Morgan, 2015). If children's real-life experience of peer conflict is that stronger consensus *doesn't* make dissenters more likely to concede, they could be justified in expecting slower decisions from larger group. They may simply not see strong consensus as a reason to expect disputes over idiosyncratic and fundamental differences to be more strictly triaged or efficiently resolved.

Task-specific decision rules, task-specific decision-speeds?

One emerging capacity that may be critical to efficient decision-making is the ability to shift between different forms of "government" depending on context and task. Among adults, the "by default" degree of consensus needed for a deliberative group to make a decision appears to be simple majority rule, in a variety of tasks and cultures — but the decision threshold groups use shifts depending on the "demonstrability" of the task (Laughlin & Ellis, 1985; Bonner et al., 2021; Boehm, 1996). In the most demonstrable tasks (e.g., involving mathematics and reasoning about physical artifacts by people with a shared conceptual system & sufficient information, motivation, and *time*), a single dissenter who can *demonstrate* the correct answer can overturn an otherwise unanimous consensus. But in "judgment" tasks (i.e., where no such demonstration is possible), groups default to various levels of majority rule. In line with this distinction, Hok et al (2025) found that 6- to 9-year-old children abide by majority rule in matters of preference (naming a rabbit), but not in matters of fact (deciding if a rabbit is actually a rabbit or a hamster). Higher-stakes decisions often require higher consensus thresholds: majority rule may suffice for deciding what type of food a group wants for dinner, but may seem inappropriate to convict someone of murder. In the demonstrability literature, mock juries typically won't

convict without a supermajority (Laughlin & Ellis, 1985; Bonner et al., 2021), and in the US, unanimity is often required by law.

To be clear, these task domains aren't intended to be bright-line differences: most tasks are multi-dimensional to some extent, and can depend as much on the group's members as the task itself (Laughlin & Ellis, 1985; Bonner et al., 2021). For instance, a patent violation for an AI-algorithm might involve both mathematical reasoning (albeit to a lesser extent than a pure arithmetic task) and moral reasoning (albeit to a lesser extent than murder trial); and which kind of reasoning matters more can depend on whether enough people in the group have sufficient expertise to follow the mathematics of the algorithm. Nevertheless, the demonstrability of a task is a useful construct for understanding how a deliberative group might approach it, and one that children appear to be sensitive to (Richardson & Keil, 2021; Hok et al., 2025).

We deliberately chose a "high demonstrability" task and emphasized egalitarian deliberation. Could changing the task domain or group composition change inferences about decisions speeds, for either children or adults? We think this is a question worth exploring in future work, but we'd still expect the people's inferences to be based on the three constraints we examined here: size, diversity, and consensus strength. Why? Because decision thresholds are simply proportions of the total size of the group. As such, dividing a group into more factions will still make it less likely that any given faction exceeds the threshold; raising or lowering the threshold will still increase or decrease the number of votes needed to exceed it; and stronger deference to stronger consensus will still be able to short-circuit endless dissent regardless of the group's diversity or total size (Albrecht, Anderson, & Vroman, 2010; Chan, Lizzeri, Suen, & Yariv, 2018; Stein, Frey, & Flache, 2024). In short, we think our drone task is relevant to a broad range of decision-making contexts, and while a task with lower demonstrability would be as well, the constraints we studied here would still apply.

Conceptual changes in decision-making procedures: balancing fairness & efficiency

That said, changes in children's reasoning about decision-making procedures in different contexts could shed light on what makes group-decision-making processes more or less efficient. Children can explicitly justify the use of different decision-making procedures in different contexts (Helwig & Kim, 1999; Hok et al., 2025), and in some contexts egalitarian deliberation may be less relevant to a decision than physical dominance, social status or alliances, or one individual's access to task-specific expertise or evidence (Mascaro & Csibra, 2012; Thomsen, et al, 2011; Heck, et al., 2021).

For instance, when groups agree to arbitrary norms (e.g., about which puppets can play where), preschoolers treat dissent as nullifying norms even with a 9-to-1 consensus in favor — although they *will* still occasionally protest if someone who *already* agreed to a norm disregards it (Schmidt, et al., 2016). And children are six-to-eight times more likely to object to an unequal distribution of resources if their group doesn't consult them first as compared to when they're given opportunity to assent (Grocke, Rossano, & Tomasello, 2018). Consulting every member of a group in advance may take less time than handling their objections, but either approach is

likely to take longer in larger groups than smaller groups regardless of initial consensus, all else being equal. And while adherence to democratic decision thresholds like majority rule may outperform other forms of government in the long run, majorities can also empower tyrants (Kawakatsu et al., 2021) or become tyrannical themselves. Any intuitive theory whose function is to help children make sense of group dynamics not only needs to take such behaviors into account — it also needs to evolve as those behaviors change across development. Our work suggests that intuitions about the time costs of consensus-based decision-making may continue to undergo conceptual change into late childhood.

A role for metacognition in managing collective speed-accuracy tradeoffs?

Critically, consensus-based decision thresholds don't speed up group decisions simply because democratic processes better capture some arbitrary eccentricity of human nature. 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 both 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 constraints 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 collective judgment so flexible even among children; other species don't invent 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), they don't use discussion to adjudicate disagreement (Domberg, Köymen, & Tomasello, 2019), and whose judgment they defer to doesn't depend on whether disagreement in the group concerns preferences, norms, or rational beliefs (Stasser, 1999; Richardson & Keil, 2022; Hok et al., 2025; Thomas et al., 2022).

Of course, stronger consensus can lead to faster decisions even without explicitly metacognitive representations of speed-accuracy tradeoffs, simply because the group is closer to its decision threshold from the outset (Conradt & List, 2009). But social dynamics in collective decisions are to some extent consequences of our beliefs about them: the less collaborators *expect* each other to concede more quickly to stronger consensus, the less dissenters may feel pressured to be judicious and efficient in picking their battles — and vice versa. 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, Alice's collaborators may treat her willingness to filibuster a growing consensus as a costly signal that her preferences, arguments, or evidence deserve more serious attention. But if they believe that Alice *expects* them to make that inference, they may instead treat her dissent as a strategic move. Formalizing quorum and

decision thresholds can make coordination easier, but they can also make our strategic negotiation tactics even more influential — swing voters can gain disproportionate power, filibusters can become vetos, and collective preferences can be overturned by manipulating agendas and gerrymandering group structures (Chan, Lizzeri, Suen, & Yariv, 2018; Levine & Plott, 1977; Pietraszewski, 2022; Stewart et al., 2019). Research on the developmental origins of commonsense reasoning about factional power and the speed-accuracy tradeoffs of collective decision-making may help us understand how that reasoning constrains social dynamics in groups. More broadly, the metacognitive capacities that make our inferences about complex social dynamics seem commonsensical may make humans especially skilled in guiding collective action (Heyes, 2016). Further research into children's reasoning about consensus strength may shed light on how we learn to manage group dynamics.

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. Recent work has suggested that cost-reward reasoning may be fundamental to commonsense psychology even in early childhood (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Liu, Ullman, Tenenbaum, Spelke, 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). But past work has typically quantified costs using physical dimensions (e.g., effort, distance traveled) or the risk of failure (Aboody, Dension, Jara-Ettinger, 2021). Time costs are more ubiquitous than physical costs: every decision takes time, regardless of whether it involves movement or a probabilistic outcome. And since time spent on a task is a matter of choice in ways physical costs can't be, time costs are more flexible and may be more difficult to interpret (Richardson & Keil, 2022). And whereas individual decision speeds are only constrained by the efficiency of the decision-maker's cognitive and biological processes and the complexity of the task itself, *collective* decision times also depend on social dynamics and our skill in managing them. For instance, Alice may interpret Bob's quick response to a complex problem as a rough estimate and want to take the time to be more precise; but if Carol and David treat Bob's answer as a precise calculation, Alice will have to decide whether challenging them is worth the time. After all, perhaps Carol and David's deference is a sign that Bob's estimates tend to be precise enough to work with, or that Bob was simply recalling a solution he'd already thought through (Richardson & Keil, 2022). Reasoning about each other's preferences for patience and precision may help collaborators collectively manage their speed-accuracy tradeoffs (Bavard et al., 2024).

Conclusions

The current work focused on reasoning about group decision speeds because regardless of how groups govern themselves, time spent deciding is a cost that has to be weighed against whatever's gained from the final decision. Like other species, humans often defer to consensus

by default (Boehm, 1996; Laughlin, 2011; Bor, et al., 2021; Claidière & Whiten, 2012; van Leeuwen et al., 2018). But while consensus judgments can be wrong, the time needed to overturn them may not be worth the attempt. Our studies suggest that 6-9 year old children infer collective decision speeds from the number of people and factions on a team, but may not have a mature understanding of consensus strength until later in development. Intuitive theories that specify when team size, diversity, or power asymmetries make some battles not worth the time may help ensure that reaching consensus isn't always akin to herding cats.

References

1. Aboody, R., Davis, I., 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*. https://doi.org/10.31234/osf.io/uymtz
2. Aboody, R., Denison, S., & Jara-Ettinger, J. (2021). Children consider the probability of random success when evaluating knowledge. *Proceedings of the Cognitive Science Society*. https://doi.org/10.31234/osf.io/a7g9t
3. 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
4. Ahl, R. E., Amir, D., & McAuliffe, K. (2024). Recalling experiences of scarcity reduces children's generosity relative to recalling abundance. *Journal of Experimental Child Psychology*, *243*, 105914. https://doi.org/10.1016/j.jecp.2024.105914
5. 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
6. 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
7. Bass, I., Bonawitz, E., Hawthorne-Madell, D., Vong, W. K., Goodman, N. D., & Gweon, H. (2022). The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development. *Cognition*, *222*, 104999. https://doi.org/10.1016/j.cognition.2021.104999
8. Bavard, S., Stuchlý, E., Konovalov, A., & Gluth, S. (2024). Humans can infer social preferences from decision speed alone. *PLOS Biology*, *22*(6), e3002686. https://doi.org/10.1371/journal.pbio.3002686
9. 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
10. Bor, A., Mazepus, H., Bokemper, S. E., & DeScioli, P. (2021). When Should the Majority Rule? Experimental Evidence for Madisonian Judgments in Five Cultures. *Journal of Experimental Political Science*, *8*(1), 41–50. https://doi.org/10.1017/XPS.2020.8
11. Brocas, I., & Carrillo, J. D. (2024). Young children build consensus in networks with local information. *Preprint*. https://isabellebrocas.org/Research/NetworkKids.pdf
12. 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
13. 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
14. 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
15. Claidière, N., & Whiten, A. (2012). Integrating the study of conformity and culture in humans and nonhuman animals. *Psychological Bulletin*, *138*(1), 126–145. https://doi.org/10.1037/a0025868
16. 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
17. 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
18. DeJesus, J. M., Venkatesh, S., & Kinzler, K. D. (2021). Young children's ability to make predictions about novel illnesses. *Child Development*, *92*(5). https://doi.org/10.1111/cdev.13655
19. Goldstone, R. L., Andrade-Lotero, E. J., Hawkins, R. D., & Roberts, M. E. (2023). The Emergence of Specialized Roles Within Groups. *Topics in Cognitive Science*, tops.12644. https://doi.org/10.1111/tops.12644

20. 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
21. 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
22. 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
23. 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
24. 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
25. 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
26. Hok, H., Gerdin, E., Zhao, X., & Shaw, A. (2025). When should the majority rule?: Children's developing intuitions about majority rules voting. *Cognition*, *260*, 106128. https://doi.org/10.1016/j.cognition.2025.106128
27. 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
28. Kawakatsu, M., Chodrow, P. S., Eikmeier, N., & Larremore, D. B. (2021). Emergence of hierarchy in networked endorsement dynamics. *Proceedings of the National Academy of Sciences*, *118*(16), e2015188118. https://doi.org/10.1073/pnas.2015188118
29. 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/
30. 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
31. Kearns, M. (2006). An Experimental Study of the Coloring Problem on Human Subject Networks. *Science*, *313*(5788), 824–827. https://doi.org/10.1126/science.1127207
32. Kearns, M. (2012). Experiments in social computation. *Communications of the ACM*, *55*(10), 56–67. https://doi.org/10.1145/2347736.2347753
33. 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
34. 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
35.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
36. 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
37. Lapidow, E., Killeen, I., & Walker, C. M. (2021). Learning to recognize uncertainty vs. recognizing uncertainty to learn: Confidence judgments and exploration decisions in preschoolers. *Developmental Science*. https://doi.org/10.1111/desc.13178
38. Laughlin, P. R. (2011). *Group Problem Solving*. Princeton University Press.
39. 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*.

40. Levine, M. E., & Plott, C. R. (1977). Agenda Influence and Its Implications. *Virginia Law Review*, *63*(4), 561. https://doi.org/10.2307/1072445
41. 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
42. 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
43. 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
44. 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
45. 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
46. 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
47. 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
48. 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
49. 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
50. Pietraszewski, D., & Shaw, A. (2015). Not by Strength Alone: Children's Conflict Expectations Follow the Logic of the Asymmetric War of Attrition. *Human Nature*, *26*(1), 44–72. https://doi.org/10.1007/s12110-015-9220-0
51. 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
52. 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
53. 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
54.Stein, J., Frey, V., & Flache, A. (2024). Talk Less to Strangers: How Homophily Can Improve Collective Decision-Making in Diverse Teams. *Journal of Artificial Societies and Social Simulation*, *27*(1), 14. https://doi.org/10.18564/jasss.5224
55. 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
56. Tomasello, M. (2021). *Becoming human: A theory of ontogeny* (First Harvard University Press paperback edition). The Belknap Press of Harvard University Press.
57. 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
58. Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and Mighty: Preverbal Infants Mentally Represent Social Dominance. *Science*, *331*(6016), 477–480. https://doi.org/10.1126/science.1199198

59. Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian Models of Conceptual Development: Learning as Building Models of the World. *Annual Review of Developmental Psychology*, 26.
60. 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

Supplemental Materials

Experiment 2: Second Task

After each trial in Experiment 2, we asked participants to predict whether or not each team in that trial would choose the 4-blade (optimal) propeller after talking.

In teams with two factions, participants tended to predict that proportionally larger faction would win, regardless of whether or not the propeller they favored was the optimal one (main text Figure 4.4; Maj.Min and SuperMaj.Maj in Table 1 below); however, the effect was not significant for the youngest children in the 2v4 team.

In the two many-faction teams, participants' predictions still appeared to depend on the proportional strength of the faction endorsing the optimal propeller. When the optimal propeller was endorsed by a plurality (6 of the 20 teammates), more than 50% of participants expected the team to choose the optimal propeller over any of the other 6 propellers. But when it was only endorsed by the 2nd largest faction (4 of the 20 teammates), participants expected the team to *not* choose it; however, the effect was not significant for the youngest children in the 4v16Div team. In other words, both children and adults *predicted* majority rule and to a lesser extent, plurality rule.

SI Table 1: binomials for final team choice predictions

	Age Group	binomial test	confidence interval	p.value
Maj.Min				
2v4	**Younger**	*20 of 50, p(Success) = 0.40*	*[0.26-0.55]*	*0.2026*
	Older	*9 of 50, p(Success) = 0.18*	*[0.09-0.31]*	*<0.001*
	Adult	*9 of 50, p(Success) = 0.18*	*[0.09-0.31]*	*<0.001*
8v4	**Younger**	*39 of 50, p(Success) = 0.78*	*[0.64-0.88]*	*<0.001*
	Older	*45 of 50, p(Success) = 0.90*	*[0.78-0.97]*	*<0.001*
	Adult	*49 of 50, p(Success) = 0.98*	*[0.89-1.00]*	*<0.001*
SuperMaj.Maj				
16v4	**Younger**	*39 of 50, p(Success) = 0.78*	*[0.64-0.88]*	*<0.001*
	Older	*44 of 50, p(Success) = 0.88*	*[0.76-0.95]*	*<0.001*
	Adult	*50 of 50, p(Success) = 1.00*	*[0.93-1.00]*	*<0.001*
6v4	**Younger**	*38 of 50, p(Success) = 0.76*	*[0.62-0.87]*	*<0.001*
	Older	*35 of 50, p(Success) = 0.70*	*[0.55-0.82]*	*0.0066*
	Adult	*48 of 50, p(Success) = 0.96*	*[0.86-1.00]*	*<0.001*
SuperMaj.PluralityDiv				
16v4	**Younger**	*34 of 50, p(Success) = 0.68*	*[0.53-0.80]*	*0.0153*
	Older	*44 of 50, p(Success) = 0.88*	*[0.76-0.95]*	*<0.001*
	Adult	*50 of 50, p(Success) = 1.00*	*[0.93-1.00]*	*<0.001*
6v14Div	**Younger**	*28 of 50, p(Success) = 0.56*	*[0.41-0.70]*	*0.4799*
	Older	*27 of 50, p(Success) = 0.54*	*[0.39-0.68]*	*0.6718*
	Adult	*32 of 50, p(Success) = 0.64*	*[0.49-0.77]*	*0.0649*
SuperMin.MinDiv				
4v16	**Younger**	*17 of 50, p(Success) = 0.34*	*[0.21-0.49]*	*0.0328*
	Older	*9 of 50, p(Success) = 0.18*	*[0.09-0.31]*	*<0.001*
	Adult	*5 of 50, p(Success) = 0.10*	*[0.03-0.22]*	*<0.001*
4v16Div	**Younger**	*20 of 50, p(Success) = 0.40*	*[0.26-0.55]*	*0.2026*
	Older	*12 of 50, p(Success) = 0.24*	*[0.13-0.38]*	*<0.001*
	Adult	*8 of 50, p(Success) = 0.16*	*[0.07-0.29]*	*<0.001*

Participant Demographics

As noted in the main text, we no longer have access to the participant-specific demographics for these experiments, as a result of transferring the lab's participant pool from Salesforce to Redcap in the fall of 2023 in order to merge it with the department database. However, we had conducted an analysis of the demographics of the participant pool in February of 2021 just before launching Experiment 1 (Experiment 1 was run from February to May of 2021, and Experiment 2 from March to June of 2022), and those demographics are very similar to the experiment-specific demographics in studies we ran in 2019 and 2020. We've included the references to those papers & text reporting the demographics below.

Paper 1's demographics (2021, data from 2019): 100+ ZIP codes from 39 U.S states, average yearly income $46,700 (imputed by ZIP code), reported races (including multi-race) 68.5% White, 12.6% declined to respond, and 18.9% identified as either Hispanic, Asian, or African-American.
REFERENCE: 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

Paper 2's demographics (2022, data from 2019-2020): 204 ZIP codes in 39 U.S. states, median yearly income $77,083 (imputed by ZIP code), reported races (including multi-race) 65% White, 9.3% Hispanic or Latino, 8.7% Asian, 8.0% declined to respond, 6.3% Black or African-American, 1% American Indian or Alaska Native, 1% Native Hawaiian or Other Pacific Islander, 0.7% Other.
REFERENCE: 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

Database demographics check on 2021.02.09 (for comparison: Exp 1 in our manuscript launched 2021.02.13, Exp 2 launched 2022.03.18): includes ZIP codes from all 50 U.S. states, median yearly income of $54,172 (imputed by ZIP code), reported race demographics (including multi-racial) 72% White (60% reporting White + no other categories), 10% Hispanic or Latino, 7% Black or African-American, 6% Asian, 2% American Indian or Alaska Native, 2% Other, 1% Native Hawaiian or Other Pacific Islander.

Responses as Likert vs. Forced-choice

We decided to use a 7-point Likert scale instead of a forced-choice judgment because we believed the confidence ratings would be informative for both children and adults for some analyses of interest. For instance, though our predictions were fairly categorical, both experiments included trials that seemed to have more "obvious" answers than others, and forced-choice judgments wouldn't have allowed us to explore those more graded inferences. Likert scales have the advantage of increasing statistical power without requiring repeat trials or larger samples. Hence, we conducted our power analysis and determined a sample size with Likert scale responses in mind. Likert scales ranging from 4-points to 21- points have been used successfully with children as young as 4 during in-person interviews by training children to use the scale by physically point at their rating. But a virtual testing environment makes physically pointing at the scale infeasible, so we elected to use a two-stage method, which we and other researchers have successfully used in the past.

One-stage, 21-point Likert scale: Bass, I., Bonawitz, E., Hawthorne-Madell, D., Vong, W. K., Goodman, N. D., & Gweon, H. (2022). The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development. *Cognition*, *222*, 104999. https://doi.org/10.1016/j.cognition.2021.104999

One-stage, 5-point Likert scale: 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

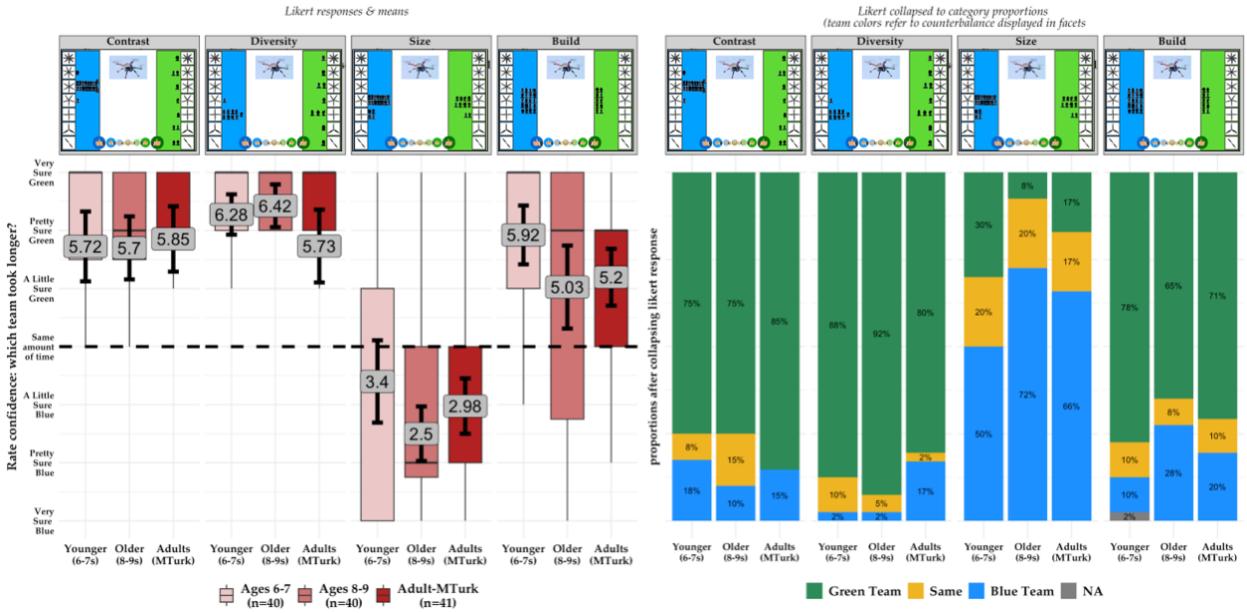
Two-stage, 5 point Likert scale: Ahl, R. E., Amir, D., & McAuliffe, K. (2024). Recalling experiences of scarcity reduces children's generosity relative to recalling abundance. *Journal of Experimental Child Psychology*, *243*, 105914. https://doi.org/10.1016/j.jecp.2024.105914

Two-stage, 4-point Likert scale: DeJesus, J. M., Venkatesh, S., & Kinzler, K. D. (2021). Young children's ability to make predictions about novel illnesses. *Child Development*, *92*(5). https://doi.org/10.1111/cdev.13655

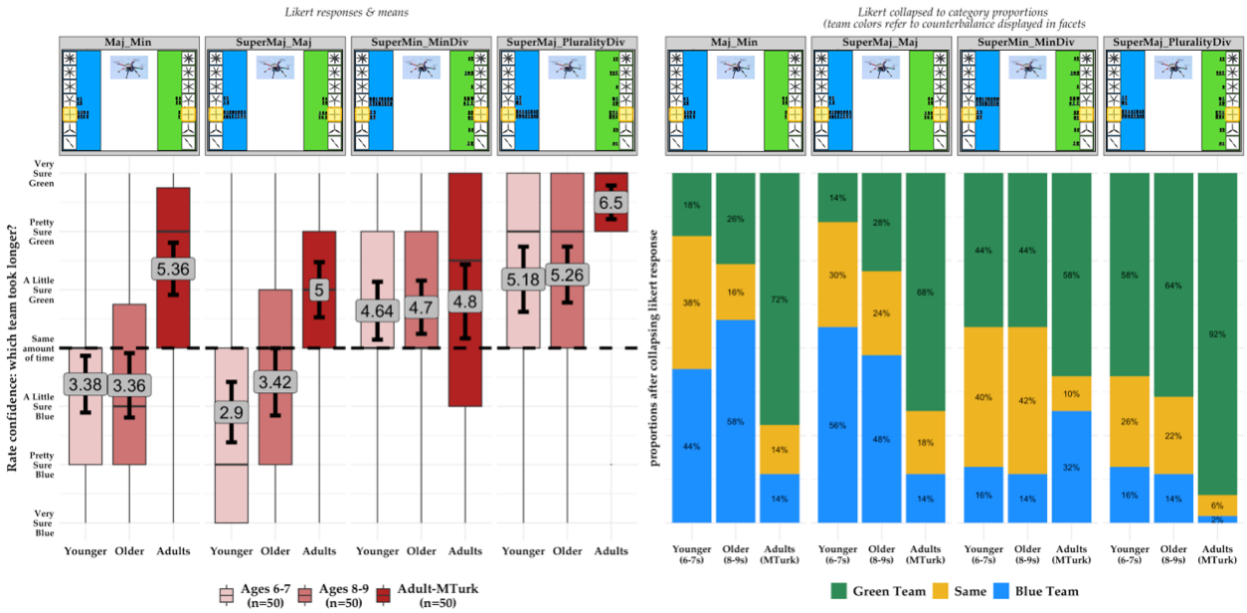
Two-stage, 3-point Likert scale: Lapidow, E., Killeen, I., & Walker, C. M. (2021). Learning to recognize uncertainty vs. recognizing uncertainty to learn: Confidence judgments and exploration decisions in preschoolers. *Developmental Science*. https://doi.org/10.1111/desc.13178

Though we designed our experiment with Likert scale responses in mind, the pattern of results doesn't change if we collapse the Likert scale to a forced-choice judgment with three categories (blueSlower-sameTime-greenSlower). The figures below show the two patterns side-by-side for each experiment. In Experiment 1, the proportion of participants making the predicted inference *at least* as large as the other two inferences combined, for both children and adults, and very few par. In Experiment 2, the predicted inference is the most common in every trial, and once again *very* few children or adults make the opposite inference. The proportion of "both teams took the same amount of time" responses is notably higher among children than among adults in every trial, but only in the *SuperMin_MinDiv* trial is the "same time" response close to as common as the predicted inference.

Experiment 1: responses as Likert vs forced choice



Experiment 2: responses as Likert vs forced choice



Power Analysis

Pilot data for Experiment 1 suggested a potential effect of counterbalance in the order of trials, such that when the *Contrast* question came first, the effects were weaker than when it came last; and, of lesser methodological concern, a potential developmental pattern for group size in the *Size* trial. Because these effects were small and were found with small pilot samples that didn't have equal numbers of participants in each counterbalance, we conducted a power analysis by simulation that modeled them as random effects. As noted in our pre-registration, we decided to treat the counterbalance effects as something to be potentially explored if they appeared in the a full sample for Experiment 1. They did not.

Because Experiment 2 involved a more challenging procedure and pilot data suggested that children's inferences may be less certain, we expected to need a larger sample than in Experiment 1. In order to determine whether increasing the sample size from n=40 per age group to n=50 would be sufficient, we conducted a power analysis that simulated responses with several effect sizes that appeared plausible based on pilot data, and looked at the coefficient estimates and significance criterion for a one-sample t.test comparing each age group to chance for each trial.