Author(s)
Emory Richardson
Hannah Hok
Frank Keil
Alex Shaw

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?

Group collaboration allows humans to accomplish things that would be impossible for individuals working alone. However, collaboration also introduces new costs. One cost is the time spent coordinating multiple perspectives to reach consensus. How do we decide if a group collaboration is worth the time? Here, we aim to examine children's intuitions about how the speed of group decision-making is influenced by *size* of the group and the *diversity of opinion* of its members.

We make three predictions.

(1) First, we predict that children will expect groups with higher-diversity of opinions to take longer than groups with lower-diversity of opinions.

(2) Additionally, we predict that children will expect a larger group to take longer to come to a decision than a smaller group with less members.

(3) Finally, we expect that group size will not be a pivotal factor across every scenario: we predict that when group size and group diversity are pitted against each other, children will expect high-diversity groups to take longer to make decisions regardless of size of group.

While we will examine developmental differences in our analysis, we expect these intuitions to develop early. Past work has suggested that while children expect physically complex tasks to take longer than physically easy tasks by age 4-5, their intuitions about time and complexity for cognitive tasks develops around age 6-7 (Leonard et al., 2019; Richardson & Keil, 2020). Thus, our primary predictions here are not of developmental change, but potentially mature intuitions even in the our youngest age group (6-7).

To test these hypotheses, we will tell children stories about two groups building drones (designated as the "Blue" Group and the "Green" Group). Children will be told that each group will have to decide how many blades the propellers should have in order to make their drone fly the best — and that drones can have anywhere between two to eight blades. However, we will say, people sometimes disagree about how many blades make drones fly the best, and that when people in a group disagree they have to talk together to make a single decision. We will then show children the two groups (Green and Blue), drawing their attention to the size of each group (e.g., the number of people in the group), as well as the distribution of "opinion" within each group (which number of blades each member thinks works the best). Across three different trials, the groups will differ only in size (10 people vs. 20 people), only in diversity (low vs. high), or in both size and diversity (20 people-low diversity vs 10 people-high diversity). Children will be asked to predict which group will take longer to discuss and decide how many blades to use.

3) Describe the key dependent variable(s) specifying how they will be measured.

Children will indicate whether they are "a little sure", "pretty sure", or "very sure" that the Blue group or Green group will take longer to decide on a 7-point Likert scale, in which the mid-point designates "the same". Children will be trained on how to use the scale before starting the experiment.

4) How many and which conditions will participants be assigned to?

There will be three conditions collected within-subjects:

Diversity – Each Group is shown to have the same number of individuals (e.g., both groups have 10 members), however the "spread" of what each individual believes is the best option is manipulated. In the Low-Diversity Group, 2 options are endorsed. In the High-Diversity group, 9 options are endorsed.

Size – Each Group is shown to have a different starting number of individuals. The Large Group will have 20 individuals whereas the Small Group will have 10 individuals. The Diversity of options will be kept identical (e.g., both groups endorse 2 options).

Diversity x Size – One of the groups is "High-Diversity x Small Group" while the other is "Low-Diversity x Large Group".

Additionally, we counterbalance the order in which these conditions appear but block together Diversity and Size, given that the Diversity-Size trials are specifically "pitting" these two variables against each other. Thus, we have four counterbalance options of trial appearance that participants will be randomly assigned to:

1. Diversity x Size; Diversity; Size
2. Diversity x Size; Size; Diversity
3. Size; Diversity; Diversity x Size

4. Diversity; Size; Diversity x Size

5) Specify exactly which analyses you will conduct to examine the main question/ hypothesis.

We test our hypotheses in two ways. We will conduct a separate linear regression on the child sample alone for each dependent measure (Div, Size, DivVsSize) using the following formula:

lm(Ct_Values~Ct_AgeYears*OrderCB, data=.)

By centering responses (Ct_Values) to the 7-point likert scale on 0 and the children's age in years (Ct_AgeYears) on the midpoint of the children's age range (7.5 years), we can make the intercept of the regression equivalent to a one sample t.test versus the scale midpoint for the mean age of the whole child sample, while the age coefficient allows us to examine developmental changes.

If the order effects are not significant, we will reduce the developmental model to simply *Ct_Values~Ct_AgeYears*.

In order to directly test beliefs about which group will take longer for each of the three measures, we will also conduct one-sample t.tests comparing each age group (Younger, Older, Adult) and measure (Div, Size, DivVsSize) to chance.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

Child participants will be excluded and replaced in case of connectivity issues with the platform, such as extreme lag, poor audio-connection, etc. Children will also be excluded and replaced if parents interfere. Finally, children will be excluded and replaced if they fail the training questions for the scale. Adult participants will be required to pass an attention check after reading the initial instructions, consisting of basic comprehension questions about the instructions. Participants who fail the attention check twice will be screened out of the study.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.

We will test 40 participants per age group (Younger, Older, Adult). This will give us sufficient power to run t.tests for each age group in addition to the tests of the entire sample of children. R scripts for the Monte Carlo simulations used to determine sample size can be found in the OSF repository.

Younger children: n=40 (20 age 6, 20 age 7)
Older children: n=40 (20 age 8, 20 age 9)
Adults (MTurk): n=40

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

Pilot results suggested a potential effect of counterbalance order, such that when the DivVsSize 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 measure. Because these effects were small and were found with small pilot samples, we modeled them as random effects in the Monte Carlo simulations, but will treat them as effects to potentially be explored in follow up studies if they appear in the experiment with a full sample.

We will also explore the relative importance of diversity and size by conducting two-sample t.tests for each age group comparing DivVsSize to Div and DivVsSize to Size.

After the experiment, we include an exploratory measure to give a more complete picture of participants' judgements: we will ask participants to judge which team will take longer to *build* the drone, given that they (A) all agree about which propeller to use, and (B) both start at the same time. Because larger teams can divide labor between more people , they may be faster to simply complete a physical task, independently of the time it takes them to coordinate consensus on *how* to accomplish the task — "many hands make light work". However, this assumes the labor can be effectively divided — if there are "too many cooks in the kitchen", larger teams may be slower than smaller teams because more time is required to allocate tasks across many team members that could be efficiently completed by one person. Intuitions about how to divide labor most efficiently may develop independently of a more basic intuition about the time costs of coordinating opinions in a group, and may warrant more direct attention in later studies, depending on the patterns we observe in the current study.