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?

In a previous experiment, we tested the predictions that children would infer that diversity (the number of unique solutions endorsed by individuals in a group) and size (the number of individuals in the group) increase group decision times, with diversity given more weight than size.

Here, we extend our findings by testing the role of factional power. Our predictions are based on a simple intuition: if *k* teammates endorse *k* different solutions, then *k-1* teammates will each have to change their opinion at least once to reach unanimous consensus. Since larger *k* will require more opinion changes, this implies longer decision-times in teams with either (A) more members, or (B) a greater diversity of endorsed solutions. However, using quorum thresholds such as plurality- or majority-rule can reduce decision times if individual voters act as factions/ blocs and are more likely to defer to factions endorsed with a greater proportional share of the overall vote: rather than needing to resolve disagreements between each individual, the group can resolve the smaller number of disagreements between factions. Moreover, to the extent that proportionally larger factions have more power over smaller factions, their influence may further increase decision speed. Thus, while diversity and size both suggest longer decision times, a large team with a clearly dominant faction (i.e., majority or plurality) may still be able to make quicker decisions than a smaller team with many factions of equal power.

However, to the extent that reasoning about the impact of power differences on group decision speed requires observers to integrate both the number of group members and the proportion endorsing each solution, reasoning about power may be more challenging than reasoning about group size or diversity alone. For instance, an agent may expect quick decisions if one faction has quorum, but nothing more granular; this could lead them to ascribe equal power to simple- and super-majorities, or weak pluralities and pluralities one vote shy of outright majority. Similarly, an agent may not expect factions to always act as cohesive blocks; this could lead them to infer that while group diversity is an important influence on decision speed, group size is more important influence than factional power (indeed, larger factions may be less

cohesive). Such patterns may be more common early in development; difficulty integrating information often leads children to rely on simpler heuristics.

We predict that all age groups will infer longer decision times in more diverse groups, as in Experiment 1 (T4). However, when size and factional power are contrasted, we predict that adults will infer longer decision times in groups whose final decisions were initially endorsed by factions with less relative power, while older children (ages 8-9) will infer longer decision times in larger groups. Because Exp 1 and pilot data for Exp 2 suggested that younger children (ages 6-7) may be less sensitive to the impact of group size despite recognizing the impact of group diversity, we make no prediction for younger children in the trials in which size and factional power are contrasted; they may show the same pattern as older children, or may not differ from chance. Thus, our primary analysis in these contrasts (T1-T2, below) will concern adults and older children.

To test these ideas, we modify the paradigm we used in previous work. As before, we show participants pairs of teams deciding on one of 7 possible solutions (namely: how many blades a propeller should have to make a remote control drone fly the best in a contest), and show them which solution to the problem *each* teammate on each team endorsed. However, in this experiment we additionally reveal that while each team ultimately endorsed the optimal solution, the optimal solution either *was* or *was not* the solution initially endorsed by the dominant faction on each team. Revealing the outcome in this way allows us to separate participants' expectations about each faction's influence on the final decision from their influence on the decision speed, and enables a more direct test of their beliefs about how diversity, size, and factional power impact decision speed. Additionally, after asking children to "pretend for a moment" that both teams had endorsed the optimal solution and guess which team took longer, we ask them to infer which solution each team had *actually* endorsed; for the accuracy measure, we expect all ages to infer majority rule in all trials — thus, they will expect the group to choose the optimal solution when favored by the initial majority, but not otherwise.

For brevity's sake, we illustrate each of the four trials below as the number of endorsements received by each of the 7 solutions on each of the 2 teams; in all four trials, both teams ultimately endorse four-blade propeller (3rd-from-left). The two teams are color-coded to facilitate children's responses in video-chat studies; color is counterbalanced between participants.

Type1: BlueTeam (0v0v***2***v4v0v0v0) vs. GreenTeam (0v0v***8***v4v0v0v0)
Type2: BlueTeam (0v0v***6***v4v0v0v0) vs. GreenTeam (0v0v***16***v4v0v0v0)
Type3: BlueTeam (2v2v***4***v6v1v3v2) vs. GreenTeam (0v0v***4***v16v0v0v0)
Type4: BlueTeam (2v2v***6***v4v1v3v2) vs. GreenTeam (0v0v***16***v4v0v0v0)

Whereas T1 & T2 contrast size and factional power, T3 & T4 contrast diversity and factional power. On our account, participants will infer that the team with the longer decision time will be as follows (given the color labels above):

Type1: Adult-Blue, Older-Green, Younger-(either Green or not sig)
Type2: Adult-Blue, Older-Green, Younger-(either Green or not sig)
Type3: Adult-(Blue or not sig), Older-(Blue or not sig), Younger-(Blue or not sig)
Type4: Adult-Blue, Older-Blue, Younger-Blue

Note that T4 contrasts a low-diversity-majority-rule group against a high-diversity-plurality-rule group, both diversity and relative power favor the same judgment, while T3 contrasts a low-diversity-minority-rule group against a high-diversity-minority-rule group; thus, we expect results overall to be strongest in T4, but weakest (if significant) in T3.

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 took longer to decide on a 7-point Likert scale, in which the mid-point designates "the same".

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

All four trials are within-subjects.

Type1:(0v0v***2***v4v0v0v0) vs. (0v0v***8***v4v0v0v0)
Type2:(0v0v***6***v4v0v0v0) vs. (0v0v***16***v4v0v0v0)
Type3:(2v2v***4***v6v1v3v2) vs. (0v0v***4***v16v0v0v0)
Type4:(2v2v***6***v4v1v3v2) vs. (0v0v***16***v4v0v0v0)

Each participant is assigned to one of four counterbalanced orders. The four OrderCBs were generated to (A) favor T1 and T2 appearing first, as we expect T4 to be most robust to any order effects if such are found, and (B) prevent T1 and T3 from appearing consecutively, as they are the most likely to interfere with each other if order effects are found (because both have minority-rule outcomes).

1. T1-T2-T3-T4
2. T2-T1-T4-T3
3. T3-T2-T1-T4
4. T4-T3-T2-T1

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

To analyze the developmental pattern for time judgments, we will use a standard one-way ANOVA procedure. Whereas in our previous experiment we expected similar results for each age group with a consistent developmental pattern, we here expect a qualitative shift from childhood to adulthood; however, because noise in younger children's responses (our previous study and pilot data revealed that they had difficulty with groups that varied in size) may lead to a non-linear developmental pattern, our primary analysis for T1-T2 will be comparing the older children (i.e., 8-9s) to adult judgments with planned contrasts, though will also compare younger children (i.e., 6-7s) to both adults and older children with post-hoc tests. In T3-T4, we expect all age groups to show a similar pattern, and so run planned contrasts for each age for the developmental pattern analysis.

To analyze qualitative preferences for each contrast Type (T1-T4), we will run one-sample t.tests comparing responses for each age group to the midpoint of the scale (i.e., 4). We are primarily interested in the adults and older children in T1-T2, but all age groups in T3-T4, as noted above

Accuracy judgments will be tested with binomial comparisons to chance (.5) for each group in each contrast (see Rscript: "22.3.4_PGS_Exp2_aovPowerAnalysis" on the OSF repository for complete predictions).

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 50 participants per age group (Younger, Older, Adult), giving us 80% power to detect a moderate-to-large effect for the test of qualitative preferences. R scripts for the Monte Carlo simulations used to determine sample size can be found in the OSF repository.

Younger children: n=50 (25 age 6, 25 age 7)

Older children: n=50 (25 age 8, 25 age 9)
Adults (MTurk): n=50

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

(1) We will ask participants to explain their answers for each time judgment, and develop a coding scheme to quantify references to group size, diversity of opinion, and factional power.

(2) As an exploratory analysis, we will examine inferences about plurality-rule for the accurate diverse group in T4.

(3) As an exploratory analysis, we will examine the importance of diversity and size cues by comparing inferences in T4 and T2, in which a 16v4 majority-accurate group is contrasted with either a 6v4 majority-accurate group (T2) or a diverse, plurality-accurate group of equal size.

(4) If even children's inferences are stronger in T4 than in T3, it may suggest some sensitivity to factional power despite their difficulty reasoning about it in T1-T2.