Supplementary Information for Freeman et al. PNAS, “Social and general intelligence improves collective action in a common pool resource system"

Jacob Freeman[a,b,1], **Jacopo A. Baggio**[c,d,1,2], **and Thomas R. Coyle**[e]

[a]Anthropology Program, Utah State University; [b]Ecology Center, Utah State University; [c]School of Politics, Security and International Affairs, University of Central Florida; [d]Sustainable Coastal Systems Cluster and National Center for Integrated Coastal Research, University of Central Florida; [e]Department of Psychology, University of Texas at San Antonio

Introduction

Sustaining collective action to govern natural resources in a globalizing world can only build upon the cognitive foundations of collective action among human groups. In this paper, we focus our analysis on the effects of general and social intelligence on the performance of small groups in a common pool resource experiment. This supplementary document provides more details about the experiment and additional analyses that buttress the results presented in the main paper. In general, we observe support for the Functional Intelligences Proposition (see also (1, 2)) that groups with with high general intelligence (g) and high social intelligence (ToM) engage in more effective and consistent collective action.

Section #1 provides a more detailed description of the experiment's design and treatments. Section #2 provides more details on how we estimate the effectiveness of collective action in the common pool resource experiment. We also discuss how we estimated general intelligence and social intelligence and describe other “control" factors that previous research suggests affects the sustainable management of common pool resources. Section #3 provides additional analyses of the regression models proposed in Fig. 1 of the main text. We provide evidence that the results are robust to adding or deleting other factors argued to impact performance in common pool resource settings. Section #4 provides details on all of the steps involved to construct and analyze the effects of general intelligence and social intelligence on the sustainability learning curves presented in Fig. 4 of the main text. Section #5 provides details on recruitment for the experiment and discusses how well the sample of participants represents the US population. Section #6 describes in more detail the short-story test and the grading of the short-story test to estimate each participant's theory of mind score. Finally, section #7 consists of the exit survey each participant completed upon completing the common pool resource experiment.

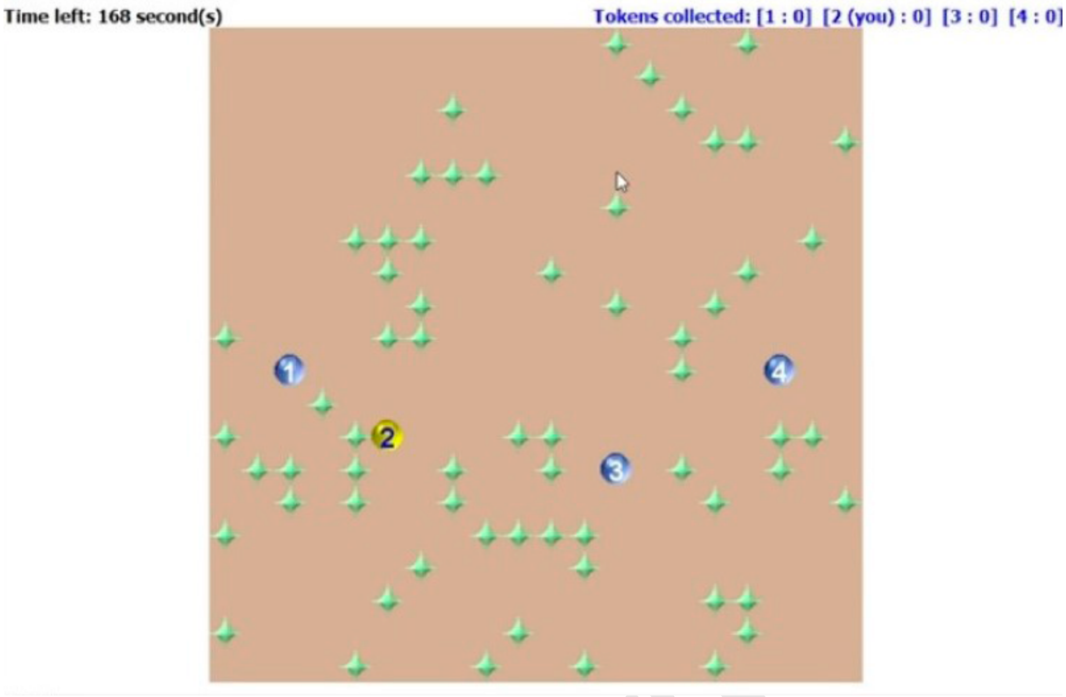
1. The Experiment

To evaluate the FIP, we used a foraging game that simulates a common pool resource system (3). The common pool resource system consists of a spatially dispersed resource (tokens) that grows according to a density dependent function. Each participant in our experiment was paid $0.02 per unit of resource harvested (tokens). Thus, individuals constantly faced the temptation to harvest as many tokens as they could, as quickly as they could, to maximize their revenue in the short-run. However, so doing always has a community cost in the experiment: the resource base may be depleted quickly and collapse.

In each experimental treatment, groups of four or eight anonymous individuals harvest tokens for 6 rounds (180 seconds each) on a $20X20$ grid. Fig. S1 provides a screen shot of each participant's view in the virtual foraging environment. Participants can see the whole grid, and thus have information on how other individuals within their own group behave. However, they have no information on other groups participating in an experimental session or how many rounds they would play the game. Typical experimental sessions included 16 to 24 participants in labs with cubical computers at Utah State University (USU) and the University of Texas at San Antonio (UTSA). Participants are assigned randomly to groups, but they can communicate with other group members before the first round of the game and between rounds thereafter by using a chat dialog box. The chat is anonymous - that is, one knows she is chatting with people in her group but does not know who they are.

At the beginning of each round, tokens fill 15% of the grid. Empty cells can generate tokens with probability $p_{tok} = p_g {*} n_{tok}/N$, where p_g is the maximum growth rate (that changes in both the resource and the group-size treatments). In the resource treatment the token-regeneration rate is 0.01 in the case of high growth rate, and 0.005 in the case of low growth rate. In the group size treatment the token regeneration rate is 0.01 in the case of groups of 8 and 0.005 in the case of groups of 4. n_{tok} = the number of neighboring cells with tokens, and $N = 8$ represents the maximum number of neighboring cells that can contain tokens. The more tokens that neighbor an empty cell, the higher the probability that the empty cell will generate a token. Conversely, if all neighboring cells are empty, no token will grow.

A. Treatments. A critical component of the experiment is that we divided game play into two three round segments under four treatments. These treatments either changed the group size or growth rate of tokens. The four treatments are coded in our data repository as the High-Low Resource treatment; Low-High Resource treatment, Group1 four-eight treatment and Group2



Time left: 168 second(s)

Tokens collected: [1 : 0] [2 (you) : 0] [3 : 0] [4 : 0]

Fig. S1. Experiment participant view example.

eight-four treatment. The High-Low Resource treatment and the Low-High Group1 treatment mimic a negative perturbation to the system, while the Low-High Resource treatment and the High-Low Group2 treatment mimic a positive perturbation to the system.

The software used for this experiment is open-source and available at http://commons.asu.edu. The software was modified in order to fit our experimental needs. In all four experimental treatments, individuals were allowed to communicate before each round of the game but were never informed about the change either in the resource growth rate nor about impending changes in group size. In sum, the four treatments were as follows:

1. Negative Perturbation–In the High-Low Resource treatment, we evaluated the effect of a negative change in the growth rate of the resource base on the ability of groups to collectively harvest tokens. In this treatment, individuals harvest tokens in rounds 1-3 with a high re-growth rate (p_g=0.01), and in rounds 4-6 with a low regrowth rate (p_g=0.005).

2. Positive Perturbation–In the Low-High Resource treatment, we evaluated the effect of a positive change in the growth rate of the resource base on the ability of groups to collectively harvest tokens. In this treatment, individuals harvest tokens in rounds 1-3 with a low growth rate (p_g=0.005), and in rounds 4-6 with a high growth rate (p_g=0.01).

3. Positive Perturbation–In the Group2 eight-four treatment, we evaluated the effect of a reduction in group size on the ability of groups to collectively harvest tokens. In this treatment, individuals harvest tokens in rounds 1-3 in a group of 8 participants, and in rounds 4-6 in a group of 4 participants. Note, we maintain a token growth rate scaled to group size. A group of 8 harvests tokens at p_g=0.01 and a group of 4 at p_g=0.005.

4. Negative Perturbation–In the Group1 four-eight treatment, we evaluated the effect of an increase in group size on the ability of groups to collectively harvest tokens. In this treatment, individuals harvest tokens in rounds 1-3 in a group of 4 participants, and in rounds 4-6 in a group of 8 participants.

B. Experiment Protocol. To facilitate the replication of this work we provide the code used to run the experiment here: https://bitbucket.org/tamnguyenthe/fip-game

Fig. S2 describes the protocol used in the experiment for all treatments. After recruitment (see next section) participants were asked to read instructions, sign a consent form, and then, if consent was given, participants were prompted to the "Read the Mind in the Eyes Test" (RMET). After the RMET they were asked to play one of the four treatments enumerated above. Once finished playing the 6 rounds of the actual common pool resource game, participants were asked to take the Short Story Test (SST) and finally an exit survey.

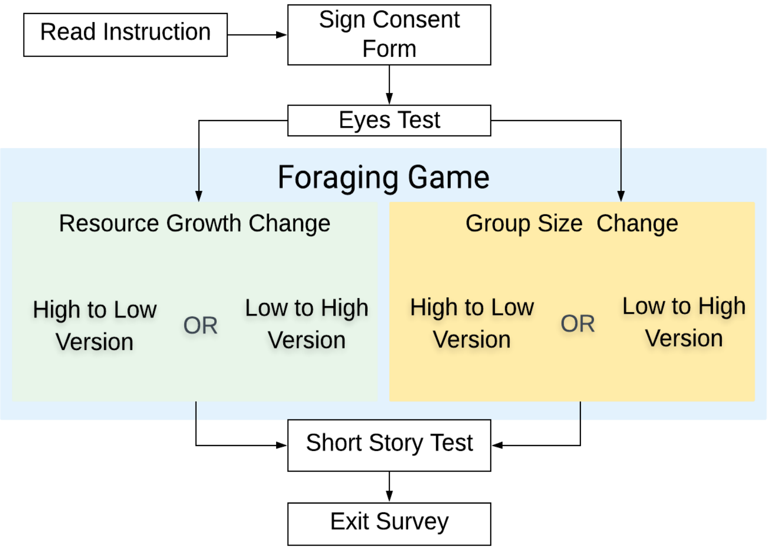


Fig. S2. Experiment protocol flow.

Once individuals were seated at their experimental stations, the following instructions where given:

Welcome to the Packman Game. You have completed the consent form, and you are now ready to participate. Please give your best effort. You will receive academic credit for participating. In addition, based on your performance, you can also earn up to $40. The experiment will take 1 to 1.5 hours, and is divided into three parts. In Part I you will complete an "eyes task". In Part II, you will play a foraging game. In Part III you will complete a "short story task" and answer a few questions on an exit survey. Please no talking during the experimental session. You must have completed the ACT or SAT and be 18 years or older to participate. Thank you for your participation!

No questions from participants were allowed during the experiment to minimize facilitator interference. The following text contains the instructions, as they appeared on screen before the common pool resource game:

In this game you will earn money for collecting tokens. The amount of money you earn depends on your decisions AND the decisions of other people in this room over the course of playing a game described below.

How to play

You will appear on the screen as a yellow dot (avatar) with other individuals who will appear as avatars. You can move by pressing the four arrow keys on your keyboard (see above Fig. S1 for an example of what participants would see on the screen during the foraging game).

You can move up, down, left, or right. You have to press a key for each and every move of your yellow dot. As you move around you can collect green diamond shaped tokens and earn two cents for each collected token. To collect a token, move your yellow dot over a green token and press the space bar. Simply moving your avatar over a token does NOT collect that token.

Between rounds of token collecting you will have 1 minute to chat via text box.

Tokens

The tokens that you collect have the potential to regenerate. After you have collected a green token, a new token can re-appear on that empty cell. The rate at which new tokens appear depends on the number of adjacent cells with tokens. The more tokens in the eight cells that surround an empty cell, the faster a new token will appear on that empty cell. Existing tokens can generate new tokens. To illustrate this, please refer to Image 1 and Image 2. The middle cell in Image 1 denoted with an X has a greater chance of regeneration than the middle cell in Image 2. When all neighboring cells are empty, there is no chance for regeneration.

Best Strategy

The chance that a token will regenerate on an empty cell increases as there are more tokens surrounding it. Therefore, you want to have as many tokens around an empty cell as possible. However, you also need empty cells to benefit from this regrowth. The best arrangement of tokens that maximizes overall regrowth is the checkerboard diagram shown below. The slower the token regrowth, the more patient you must be in order for a token to reappear after harvest.

2. Measuring Dependent and Independent Variables

We use four dependent variables to measure how well groups perform in the common pool resource system. The main dependent variable is the proportion of time ($Time$) per group per round that groups leave tokens in the foraging environment. For example, if a group harvested all of the tokens in 10 seconds, they would have left tokens in the commons for 10 $second/180$ $seconds$ for a proportion of 0.05 (poor collective action).The second variable is a normalized measure of token harvest ($Tokens$) that assesses the proportion of the maximum possible number of tokens per group (or how close a group is to harvesting a simulated maximum of harvested tokens per person). We use these variable, as discussed in the main paper, to estimate group performance on average ($Mtime$ & $Mtokens$), performance in round one ($R1$ $Time$ & $R1$ $Tokens$) and performance over all six rounds. In addition to $Time$ and $Tokens$, we calculated the coefficient of variation for $Time$ and $Tokens$ over successive rounds of game play to measure how consistently groups performed over time.

The main variable we use to measure the extent of collective action to sustain the resource is the $Time$ per round that groups leave tokens on the screen. This is our preferred measure of collective action because participants must coordinate their harvest in order to leave tokens on the screen throughout a given round. If participants do not attend to the joint goal of waiting until near the end of a round to harvest, they will deplete the resource and cause it to collapse early in a round. The growth function of the resource is density dependent, as specified above, and early in a round the resource is very sensitive to harvest. Thus, even if one or two individuals decide to harvest early in the round, this can lead the resource to collapse well before the end of a round. In addition, the $Time$ that groups leave tokens on a screen is comparable across treatments without normalizing for differences in the growth rate of the resource or group size.

We analyze four dependent $Time$ variables. The first is simply the mean time over all six rounds that each group, regardless of group size or growth rate, leaves tokens on the screen per round ($Mtime$). $Mtime$ =0.819 in the "negative" treatments with a standard deviation of 0.139, and $Mtime$ =0.826 in the "positive" treatments with a standard deviation of 0.122. The second variable is the coefficient of variation CV $Time$ over rounds 2-6. Higher values indicate less consistent collective action and lower values indicate more consistent collective action over rounds. We chose to use this metric of consistency because the coefficient of variation allows us to compare across treatment types and groups. The standard deviation should technically be understood in relationship to a group's mean performance and comparisons across groups, especially groups from different treatments, are less meaningful due to treatment induced differences in mean performance. Finally, we use $Time$ in round one to estimate how well groups perform in the most novel situation in the experiment (round one) and $Time$ over all six rounds as the dependent variable to analyze sustainability learning curves. This allows us to study how collective action to govern the resource changes over six rounds for different treatments.

We analyze three dependent $Tokens$ variables. One cannot simply compare the number of tokens harvested per group member per round across treatments. This is because the growth rate of the resource and group size varies across treatments and within treatments. To develop a comparable measure of tokens, Baggio and colleagues (2), ran simulations of the foraging game to determine the distribution of the maximum number of tokens harvested per simulated round, given the two different growth rates of the tokens used in the experiment. The maximum tokens ($Max_{T}okens$) is then used to determine how groups performed in a specific round. Formally: $Tokens = Actual_{t}okens/Max_{T}okens$ where $Actual_{T}okens$ is the number of tokens actually harvested by a group in a specific round. This gives us a ratio variable that typically scales between 0 and 1 (though the value can go slightly above 1 as the simulations and real game have a stochastic element); 1 indicates maximum harvest of tokens and 0 indicates that no one in a group harvested any tokens. $Tokens$, thus, estimates the ability of groups to harvest compared to the statistical maximum identified from 1000 simulations. The variable $MTokens$ measures the mean of normalized token harvest per group over all six round. And CV $Tokens$ measures the coefficient of variation of normalized token harvest per group over rounds 2-6. As above, $R1$ $Tokens$ measures the ability of groups to maximize token harvest in round one of the game.

An important note about the two dependent variables: $Time$ and $Tokens$ are positively related, but not perfectly as they measure different aspects of performance in the commons. $Time$ measures the ability to coordinate harvest to leave tokens in the commons (i.e., sustain the resource). $Tokens$ measures the ability of groups to maximize the number of tokens collected. Effective cooperation (higher values of $Time$) is necessary to maximize $Tokens$, but not sufficient. For example, a group could leave tokens on the screen at the end of a round–effectively sustaining the resource–but failing to maximize the harvest of tokens (each round is self contained and thus no advantage exists to leave tokens unharvested). The correlation between $Mtime$ and $Mtokens$ is r=0.49, p<0.05.

G, general intelligence, was estimated by official ACT/SAT scores released by the experiment participants. ACT/SAT scores correlate with IQ scores ($r = 0.86$) (4–6)). Given that participants could report either SAT or ACT scores we used the College Board equivalence tables to transform SAT into ACT scores (7) https://collegereadiness.collegeboard.org/pdf/higher-ed-brief-sat-concordance.pdf.

ToM, theory of mind, was estimated via a short-story test first proposed by Dodell-Feder and colleagues (8). This short-story test assigns values to an individual social reasoning ability (8). The test was taken by all participant in the experiment after the game was played (see also S2). In short, the short story test estimates social-cognitive theory of mind, which is the ability of individuals to assess and understand others' attention and intentions, and consequently, allows individuals to assess and plan joint courses of action. Social cognitive ToM allows individuals to reduce the behavioral uncertainty of others (see section #6 for more on this test). Social-cognitive ToM is conceptually distinct from social-perceptual ToM (9), which refers to an individual's ability to make a snap judgment about others' emotions from their eyes.

Following Baggio and colleagues (2), in order to assess cognitive functional diversity at the group level we use average g and minimum ToM. This is because g and ToM serve different functions. G relates to understanding the dynamics of a system and this can be thought of a cumulative task (i.e., 'I understand X, and you understand Y', together one could argue, we understand X+Y > X and X+Y > Y) (10). On the other hand, ToM refers to social interaction and hence the modeling ability is akin to a conjunctive task where the minimum level of ToM determines the "tenure" of the interaction within a group (10, 11). Groups with a high level of functional diversity, thus, have a high average g and a high minimum ToM score.

In addition to g and ToM, we control for variables found important within the literature and in previous work (2, 3, 12–16). More precisely we control for gender (% females within a group) as well as for ethnic and religious diversity, calculated as $-\Sigma(C_i * log(C_i))$ where $C_i = N_i/N$ represents the fraction of individuals of ethnicity or religion i within a group of N individuals. Ethnicity, religious affiliation and gender are calculated from the survey that participants completed at the end of the experiment session (see figure S2, as well as section #7).

Finally, we control for whether a treatment included a positive or negative shock. This is a simple dummy variable called $NegPert$. 0=negative shock to the system (either an increase in group size or decline in resource growth rate; 1=positive shock to the system (either a decrease in group size or an increase in resource growth rate). We use these variables to check the robustness of the effects of intelligence variables on collective action and the ability to maximize harvest in the analyses below. We control for this variable in all regressions. We also use a dummy variable to control for whether shocks were social or ecological ($SocPert$). Again this is a dummy variable where 0=ecological perturbation and 1=social perturbation (change in group size).

3. Mean and Consistent Performance Analysis

In this section, we provide supporting analyses for the results presented in Fig. 2. of the main text. Fig. 2A in the main text illustrates the effects of g and ToM on the average proportion of time that groups left tokens in the commons over all six rounds, $Mtime$ (i.e., the effectiveness of collective action to sustain the resource). To analyze the effects of g and ToM on $Mtime$, we used the following gls regression:

$$Mtime = \beta_a + a_1 g + a_2 ToM + a_3 NegPert + a_i k_i + \epsilon. \quad [1]$$

Where β_a is an unknown constant; g is general intelligence, ToM is theory of mind; $NegPert$ is a dummy variable that codes for either a positive or negative perturbation in round four; k_i is a vector of control variables; and ϵ is the error term. Note, we also used a glm model with a quasibimodal link function to regress $Mtime$ on the predictor variables above. Qualitatively, the results are exactly the same. This means that the direction of effects and consistency of those effects as well as significance of those effects does not change, though the exact values and interpretations of the coefficients would change with this alternative regression method.

Table S1. General least squares regression coefficients of predictor variables regressed on $Mtime$.

Model Coefficients

Predictors	Intell. Model	Control1	Control2	Control3	Control4	Control5	Control6	Control7	Control8	Control9	Control10
g	0.013***	0.013***	0.012***	0.011**	0.011**	0.013***	0.012**	0.012**	0.011**	0.012**	0.011**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
ToM	0.015***	0.014***	0.015***	0.015***	0.014***	0.014***	0.014***	0.013**	0.14***	0.013**	0.014**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$NegPert$	0.014	0.014	0.014	0.013	0.011	0.015	0.013	0.014	0.014	0.012	0.012
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
$SocPert$		-0.023				-0.026	-0.01	-0.021	-0.021	-0.024	-0.021
		(0.023)				(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)
$Gender$			0.02			0.03			0.041	0.04	0.043
			(0.049)			(0.05)			(0.05)	(0.05)	(0.05)
$ReligousDiv$				-0.047*			-0.045*		-0.047*		-0.034
				(0.026)			(0.026)		(0.026)		(0.03)
$EthnicDiv$					-0.03			-0.036		-0.038	-0.023
					(0.024)			(0.024)		(0.024)	(0.027)
Constant	0.42***	0.42***	0.40***	0.49***	0.48***	0.40***	0.49***	0.47***	0.47***	0.45***	0.48***

Model Fit

N	122	122	122	122	122	122	122	122	122	122	122
AIC	-129.24	-122.59	-123.31	-125.12	-124.14	-116.88	-118.05	-117.28	-112.60	-111.79	-105.96
Log likelihood	69.62	67.29	67.65	68.56	68.07	65.44	66.02	65.64	64.30	63.89	61.98

Note: Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level.

Table S1 illustrates that the results presented in the main text are robust to adding and deleting control variables. The first column of the table presents the coefficients and standard errors of the main Intelligences Model that regresses g, ToM and $NegPert$ on $Mtime$. The results of this model are presented in Fig. 2A of the main text. The subsequent columns illustrate coefficients, standard errors, significance values and model fits for ten control models that assess the robustness of the effects of ToM and g and significance of these variables to adding and deleting control variables argued to affect collective action by previous researchers. The main result relevant to our main argument illustrated by Table S1 is that the direction and significance of the effects of g and ToM are consistent across all models.

To analyze the effects of g and ToM on $Mtokens$, we used the following gls regression:

$$Mtokens = \beta_a + a_1 g + a_2 ToM + a_3 NegPert + a_i k_i + \epsilon \tag{2}$$

Where β_a is an unknown constant; g is general intelligence, ToM is theory of mind; $NegPert$ is a dummy variable that codes for either a positive or negative perturbation in round four; k_i is a vector of control variables; and ϵ is the error term.

Table S2. General least squares regression coefficients of predictor variables regressed on $Mtokens$.

Model Coefficients

Predictors	Intell. Model	Control1	Control2	Control3	Control4	Control5	Control6	Control7	Control8	Control9	Control10
g	0.013***	0.011***	0.014***	0.012***	0.012***	0.011***	0.008**	0.009**	0.008**	0.009***	0.008**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ToM	0.001	0.004	0.001	0.002	0.-0002	0.004	0.005	0.002	0.005	0.002	0.004
	(0.004)	(0.004)	(0.004)	(0.004)	(0.04)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$NegPert$	-0.009	-0.01	-0.009	-0.01	-0.013	-0.015	-0.012	-0.014	-0.012	-0.015	-0.015
	(0.017)	(0.016)	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
$SocPert$		0.067***				0.072***	0.074***	0.071***	0.078***	0.075***	0.078***
		(0.017)				(0.017)	(0.017)	(0.16)	(0.17)	(0.016)	(0.016)
$Gender$			0.043			-0.067*			-0.05	-0.056	0.052
			(0.039)			(0.037)			(0.036)	(0.036)	(0.035)
$ReligousDiv$				-0.053**			-0.064***		-0.061***		-0.039*
				(0.02)			(0.019)		(0.019)		(0.021)
$EthnicDiv$					-0.057***			-0.062***		-0.06***	-0.042**
					(0.018)			(0.017)		(0.017)	(0.019)
Constant	0.3***	0.3***	0.32***	0.39***	0.39***	0.34***	0.41***	0.40***	0.43***	0.43***	0.46***

Model Fit

N	122	122	122	122	122	122	122	122	122	122	122
AIC	-182.39	-188.08	-176.98	-181.01	-183.30	-184.60	-190.81	-191.96	-186.45	-187.57	-182.95
Log likelihood	96.16	100.04	94.49	96.50	97.65	99.30	102.40	102.98	101.22	101.78	100.47

Note: Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level.

Table S2 illustrates that the results presented in the main text are robust to adding and deleting control variables. As above, the first column of the table presents the coefficients and standard errors of the main Intelligences Model that regresses g, ToM and $NegPert$ on $Mtokens$. The results of this model are presented in Fig. 2B of the main text. The subsequent columns illustrate coefficients, standard errors, significance values and model fits for ten control models that assess the robustness of the effects of ToM and g and significance of these variables to adding and deleting control variables argued to affect collective action by previous researchers. The main result illustrated by Table S2 is that the direction and significance of the effect of g is consistent across all models. Interestingly, ToM has a weak and non-significant effect across all models.

Note also, unlike our collective action dependent variable, whether groups faced a social or ecological perturbation has a positive and significant effect on the ability of groups to maximize tokens. Groups who faced a social perturbation, on average, were closer to maximizing harvest than groups who faced an ecological perturbation. Finally, religious and ethnic diversity consistently have negative and significant effects on the ability of groups to maximize tokens.

A. Path models. To model the paths described by Fig. 1B and Fig. 2C and Fig. 1C and 2D in the main body of the paper, we used the following simultaneous regression equations:

$$R1time = \beta_a + a_1 g + a_2 ToM + a_3 NegPert + a_i k_i + \epsilon \tag{3}$$

$$CVTime = \beta_b + b_1 R1time + b_2 g + b_3 ToM + b_4 NegPert + b_i k_i + \epsilon. \tag{4}$$

and

$$R1tokens = \beta_a + a_1 g + a_2 ToM + a_3 NegPert + a_i k_i + \epsilon \tag{5}$$

$$CVTokens = \beta_b + b_1 R1tokens + b_2 g + b_3 ToM + b_4 NegPert + b_i k_i + \epsilon. \tag{6}$$

Where β_i is a constant in a given regression; $R1$ $Time$ is the time groups spend with tokens on a screen in round one and CV $Time$ is the coefficient of variation of time with tokens on screen in rounds two-to-six; $R1$ $tokens$ is number of tokens that a group collected in round one; CV $Tokens$ is the coefficient of variation of tokens collected between rounds two and six. G is general intelligence; ToM is theory of mind; $NegPert$ is a dummy variable indicating whether there was a positive or negative perturbation on the system; and k_i is a given control variable i including religious diversity, ethnic diversity and % of female composing a group ($Gender$) and whether the perturbation related to resources or group size ($SocPert$). To estimate these regression equations, we used Stata Structural equation modeller and replicated the results with the Lavaan package in R. We ran simultaneous regressions and used bootsrapped standard errors with 1000 draws for ten total regression models. The first model, we call the Intelligence Model, does not include the four control variables used: religious and ethnic diversity, gender composition, and whether the perturbation was related to the resources or the group size. The remaining eight control models include various combinations of these control variables to check the robustness of the effects of g and ToM.

Tables S3, S4, S5 illustrate the standardized total, indirect and direct effects of predictor variables on the ability of groups to sustain resources ($Time$) in round one and the consistency of resource sustainability ($CV\ Time$). Both g and ToM increase the ability of groups to sustain resources in round one. These effects are positive and significant across all control models.

Both g and ToM consistently, on average, reduce the variation of performance (i.e, reduce $CV\ Time$) through round one. The total and indirect effects of ToM are always significant across all control models. However, only the indirect effect of g is significant across all control models. The direction of the total effect of g on $CV\ Time$ is negative across all control models, but only significant in the Intelligence Model.

Round one also has a negative and significant effect on $CV\ Time$ across all control models. This is consistent with previous research in which group and personal characteristics may influence round one and round one then influences a group's performance in subsequent rounds due to reputation and reciprocity effects (12, 13, 17).

Tables (S6, S7, S8) illustrate the standardized total, indirect and direct effects of predictor variables on the ability of of groups to maximize tokens. A noted in the main paper, only the treatment type $NegPert$ consistently has a strong, negative, and significant effect on the ability of groups to maximize tokens in round one. G and ToM have very weak non-significant effects on the ability of groups to maximize the production of tokens in round one.

Again, as noted in the main paper, round one has a moderate, positive and significant effect $CV\ Tokens$. The better groups are a t maximizing token harvest in round one, the more variable their performance in rounds two through six. G and ToM also both indirectly increase variation in token maximization ($Tokens$) through their effect on round one; however, this effect is, again, weak and not significant in any regression model.

Finally, while g directly increases variation between rounds, ToM has a dampening effect on such variation. These effects are always significant for ToM and significant for g when the type of perturbation (whether it is centered on resource or on group size -$SocialPert$-) is not taken into account. Further, religious diversity increases variability in the ability of groups to consistently maximize tokens. Once again, the effect of religious diversity is significant only when the type of perturbation (whether social or resource centered) is not included in a model.

To summarize, g and ToM have a significant and positive effect on the ability of groups to sustain resources in round one. That is, groups with higher g and ToM are more able to sustain resources from the start (R1 results) and do so consistently by reducing variation in outcomes (see negative effects of g and ToM on $CV\ Time$ and $R1Time$ in Tables S3, S4 and S5) both directly and indirectly. On the other hand, both g and ToM have weak and insignificant effects on the ability of groups to maximize resource harvest in round one and consistently across rounds.

Table S3. Standardized Total Effects of g and ToM and control variables on collective action for resource sustainability ($Time$)

Predictors	Intell. Model	Ctr1	Ctr2	Ctr3	Ctr4	Ctr5	Ctr6	Ctr7	Ctr8
$R1Time$	-0.317***	-0.303***	-0.306***	-0.312***	-0.306***	-0.305***	-0.305***	-0.306***	-0.305***
	(0.046)	(0.046)	(0.046)	(0.047)	(0.049)	(0.048)	(0.047)	(0.047)	(0.046)
g	-0.150*	-0.111	-0.103	-0.096	-0.102	-0.097	-0.098	-0.099	-0.113
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
ToM	-0.184**	-0.197**	-0.176**	-0.179**	-0.173**	-0.174**	-0.176**	-0.176**	-0.161*
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
$NegPert$	-0.045	-0.040	-0.033	-0.035	-0.036	-0.035	-0.036	-0.036	-0.037
	(0.034)	(0.033)	(0.034)	(0.032)	(0.033)	(0.034)	(0.034)	(0.032)	(0.033)
$ReligiousDiv$		0.191**	0.132	0.140	0.138	0.159	0.164	0.163	0.155
		(0.039)	(0.047)	(0.047)	(0.045)	(0.047)	(0.046)	(0.046)	(0.048)
$EthnicDiv$			0.119	0.126	0.126	0.125	0.114	0.113	0.113
			(0.043)	(0.045)	(0.045)	(0.046)	(0.045)	(0.046)	(0.048)
$Gender$				-0.103	-0.107	-0.106	-0.106	-0.089	-0.098
				(0.076)	(0.081)	(0.078)	(0.075)	(0.088)	(0.083)
$SocialPert$					0.030	0.030	0.030	0.030	0.096
					(0.036)	(0.036)	(0.035)	(0.034)	(0.037)
Dep Var: $R1Time$									
g	0.247***	0.247***	0.247***	0.247***	0.247***	0.233**	0.236**	0.240**	0.285***
	(0.013)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.013)
ToM	0.198**	0.198**	0.198**	0.198**	0.198**	0.203**	0.209**	0.207**	0.161*
	(0.014)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
$NegativePert$	-0.060	-0.060	-0.060	-0.060	-0.060	-0.061	-0.059	-0.060	-0.055
	(0.060)	(0.059)	(0.061)	(0.058)	(0.062)	(0.061)	(0.062)	(0.062)	(0.061)
$ReligiousDiv$						-0.071	-0.090	-0.085	-0.059
						(0.069)	(0.079)	(0.082)	(0.082)
$EthnicDiv$							0.038	0.041	0.041
							(0.069)	(0.072)	(0.074)
$Gender$								-0.057	-0.026
								(0.132)	(0.132)
$SocialPert$									-0.217**
									(0.061)
N	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000
AIC	1278.968	1410.157	1530.372	1510.247	1673.265	1674.585	1676.442	1677.998	1673.847
RSE	0.000	0.000	0.000	0.000	0.084	0.101	0.137	0.205	0.000
CFI	1.000	1.000	1.000	1.000	0.911	0.903	0.880	0.866	1.000

Note: Bootstrapped Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level. RSE=Residual square error, AIC = Akaike Information Criterion, CFI = Comparative Fit Index.

Table S4. Standardized Indirect Effects of g and ToM and control variables on collective action for resource sustainability ($Time$)

Predictors	Intell. Model	Ctr1	Ctr2	Ctr3	Ctr4	Ctr5	Ctr6	Ctr7	Ctr8
g	-0.078**	-0.075**	-0.076**	-0.077**	-0.076**	-0.071**	-0.072**	-0.073**	-0.087**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ToM	-0.063*	-0.060*	-0.061*	-0.062*	-0.061**	-0.062**	-0.064**	-0.063**	-0.049**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$NegPert$	0.019	0.018	0.018	0.019	0.018	0.019	0.018	0.018	0.017
	(0.011)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)
$ReligiousDiv$						0.022	0.027	0.026	0.018
						(0.012)	(0.014)	(0.014)	(0.014)
$EthnicDiv$							-0.012	-0.013	-0.012
							(0.012)	(0.012)	(0.012)
$Gender$								0.017	0.008
								(0.022)	(0.022)
$SocialPert$									0.066
									(0.013)
N	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000
AIC	1278.968	1410.157	1530.372	1510.247	1673.265	1674.585	1676.442	1677.998	1673.847
RSE	0.000	0.000	0.000	0.000	0.084	0.101	0.137	0.205	0.000
CFI	1.000	1.000	1.000	1.000	0.911	0.903	0.880	0.866	1.000

Note: Bootstrapped Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level. RSE=Residual square error, AIC = Akaike Information Criterion, CFI = Comparative Fit Index.

Table S5. Standardized Direct Effects of g and ToM and control variables on collective action for resource sustainability ($Time$)

Predictors	Intell. Model	Ctr1	Ctr2	Ctr3	Ctr4	Ctr5	Ctr6	Ctr7	Ctr8
$R1Time$	-0.317***	-0.303***	-0.306***	-0.312***	-0.306***	-0.305***	-0.305***	-0.306***	-0.305***
	(0.046)	(0.046)	(0.046)	(0.047)	(0.049)	(0.048)	(0.047)	(0.047)	(0.046)
g	-0.071	-0.036	-0.027	-0.018	-0.026	-0.026	-0.026	-0.026	-0.026
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
ToM	-0.121	-0.137*	-0.115	-0.118	-0.113	-0.112	-0.112	-0.113	-0.112
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
$NegativePert$	-0.064	-0.059	-0.052	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054
	(0.033)	(0.032)	(0.033)	(0.031)	(0.032)	(0.033)	(0.033)	(0.032)	(0.032)
$ReligiousDiv$		0.191**	0.132	0.140	0.138	0.137	0.137	0.137	0.137
		(0.039)	(0.047)	(0.047)	(0.045)	(0.047)	(0.045)	(0.046)	(0.047)
$EthnicDiv$			0.119	0.126	0.126	0.125	0.125	0.126	0.125
			(0.043)	(0.045)	(0.045)	(0.046)	(0.044)	(0.045)	(0.046)
$Gender$				-0.103	-0.107	-0.106	-0.106	-0.106	-0.106
				(0.076)	(0.081)	(0.078)	(0.075)	(0.083)	(0.079)
$SocialPert$					0.030	0.030	0.030	0.030	0.030
					(0.036)	(0.036)	(0.035)	(0.034)	(0.036)
$Constant$	2.531***	1.824*	1.635*	1.835*	1.810*	1.803*	1.805*	1.808*	1.804*
	(0.168)	(0.185)	(0.179)	(0.184)	(0.187)	(0.179)	(0.190)	(0.199)	(0.188)
Dep Var:$R1Time$									
g	0.247***	0.247***	0.247***	0.247***	0.247***	0.233**	0.236**	0.240**	0.285***
	(0.013)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.013)
ToM	0.198**	0.198**	0.198**	0.198**	0.198**	0.203**	0.209**	0.207**	0.161*
	(0.014)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
$NegativePert$	-0.060	-0.060	-0.060	-0.060	-0.060	-0.061	-0.059	-0.060	-0.055
	(0.060)	(0.059)	(0.061)	(0.058)	(0.062)	(0.061)	(0.062)	(0.062)	(0.061)
$ReligiousDiv$						-0.071	-0.090	-0.085	-0.059
						(0.069)	(0.079)	(0.082)	(0.082)
$EthnicDiv$							0.038	0.041	0.041
							(0.069)	(0.072)	(0.074)
$Gender$								-0.057	-0.026
								(0.132)	(0.132)
$SocialPert$									-0.217**
									(0.061)
$Constant$	-1.313	-1.313	-1.313	-1.313	-1.313	-1.040	-1.098	-0.982	-0.734
	(0.332)	(0.338)	(0.327)	(0.325)	(0.342)	(0.365)	(0.381)	(0.377)	(0.385)
N	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000
AIC	1278.968	1410.157	1530.372	1510.247	1673.265	1674.585	1676.442	1677.998	1673.847
RSE	0.000	0.000	0.000	0.000	0.084	0.101	0.137	0.205	0.000
CFI	1.000	1.000	1.000	1.000	0.911	0.903	0.880	0.866	1.000

Note: Bootstrapped Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level. RSE=Residual square error, AIC = Akaike Information Criterion, CFI = Comparative Fit Index.

Table S6. Standardized Total Effects of g and ToM and control variables on resource maximization ($Tokens$)

Predictors	Intell. Model	Ctr1	Ctr2	Ctr3	Ctr4	Ctr5	Ctr6	Ctr7	Ctr8
$R1Tokens$	0.207**	0.197*	0.204**	0.193*	0.012	0.012	0.012	0.012	0.012
	(0.091)	(0.092)	(0.091)	(0.093)	(0.082)	(0.081)	(0.079)	(0.083)	(0.078)
g	0.187**	0.240***	0.247***	0.242***	0.094	0.094	0.094	0.094	0.093
	(0.007)	(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
ToM	-0.251***	-0.268***	-0.249***	-0.246***	-0.097*	-0.097*	-0.097*	-0.097*	-0.097*
	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
$NegPert$	-0.031	-0.024	-0.017	-0.016	-0.034	-0.034	-0.034	-0.034	-0.034
	(0.030)	(0.031)	(0.028)	(0.028)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
$ReligiousDiv$		0.256***	0.201**	0.195*	0.121	0.121	0.121	0.121	0.121
		(0.035)	(0.042)	(0.042)	(0.032)	(0.033)	(0.031)	(0.030)	(0.032)
$EthnicDiv$			0.111	0.105	0.094	0.094	0.093	0.093	0.093
			(0.041)	(0.039)	(0.029)	(0.028)	(0.028)	(0.028)	(0.029)
$Gender$				0.086	0.005	0.005	0.005	0.006	0.005
				(0.065)	(0.062)	(0.059)	(0.062)	(0.057)	(0.061)
$SocialPert$					0.705***	0.705***	0.705***	0.705***	0.707***
					(0.022)	(0.022)	(0.021)	(0.022)	(0.022)
Dep Var: $R1Tokens$									
g	0.045	0.045	0.045	0.045	0.045	0.051	0.046	0.039	-0.002
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
ToM	0.057	0.057	0.057	0.057	0.057	0.055	0.043	0.047	0.089
	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
$NegativePert$	-0.528***	-0.528***	-0.528***	-0.528***	-0.528***	-0.527***	-0.531***	-0.530***	-0.535***
	(0.029)	(0.030)	(0.030)	(0.030)	(0.028)	(0.030)	(0.031)	(0.030)	(0.031)
$ReligiousDiv$						0.029	0.062	0.054	0.031
						(0.036)	(0.043)	(0.045)	(0.044)
$EthnicDiv$							-0.068	-0.074	-0.073
							(0.039)	(0.039)	(0.038)
$Gender$								0.096	0.069
								(0.072)	(0.074)
$SocialPert$									0.198**
									(0.032)
N	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000
AIC	1099.116	1226.533	1346.978	1327.367	1400.472	1402.330	1403.775	1404.191	1399.839
RSE	0.000	0.000	0.000	0.000	0.097	0.122	0.156	0.209	0.000
CFI	1.000	1.000	1.000	1.000	0.969	0.964	0.961	0.965	1.000

Note: Bootstrapped Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level. RSE=Residual square error, AIC = Akaike Information Criterion, CFI = Comparative Fit Index.

Table S7. Standardized Indirect Effects of g and ToM and control variables on resource maximization ($Tokens$)

Predictors	Intell. Model	Ctr1	Ctr2	Ctr3	Ctr4	Ctr5	Ctr6	Ctr7	Ctr8
g	0.009	0.009	0.009	0.009	0.001	0.001	0.001	0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ToM	0.012	0.011	0.012	0.011	0.001	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
$NegPert$	-0.109	-0.104	-0.108*	-0.102*	-0.006	-0.006	-0.006	-0.006	-0.006
	(0.019)	(0.019)	(0.019)	(0.019)	(0.017)	(0.016)	(0.016)	(0.017)	(0.016)
$ReligiousDiv$						0.000	0.001	0.001	0.000
						(0.001)	(0.002)	(0.002)	(0.001)
$EthnicDiv$							-0.001	-0.001	-0.001
							(0.002)	(0.003)	(0.002)
$Gender$								0.001	0.001
								(0.007)	(0.004)
$SocialPert$									0.002
									(0.006)
N	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000
AIC	1099.116	1226.533	1346.978	1327.367	1400.472	1402.330	1403.775	1404.191	1399.839
RSE	0.000	0.000	0.000	0.000	0.097	0.122	0.156	0.209	0.000
CFI	1.000	1.000	1.000	1.000	0.969	0.964	0.961	0.965	1.000

Note: Bootstrapped Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level. RSE=Residual square error, AIC = Akaike Information Criterion, CFI = Comparative Fit Index.

Table S8. Standardized Direct Effects of g and ToM and control variables on resource maximization ($Tokens$)

Predictors	Intell. Model	Ctr1	Ctr2	Ctr3	Ctr4	Ctr5	Ctr6	Ctr7	Ctr8
$R1Tokens$	0.207**	0.197*	0.204**	0.193*	0.012	0.012	0.012	0.012	0.012
	(0.091)	(0.092)	(0.091)	(0.093)	(0.082)	(0.081)	(0.079)	(0.083)	(0.078)
g	0.178**	0.231**	0.238***	0.233***	0.093	0.093	0.093	0.093	0.093
	(0.006)	(0.007)	(0.006)	(0.007)	(0.005)	(0.005)	(0.004)	(0.005)	(0.004)
ToM	-0.263***	-0.279***	-0.260***	-0.257***	-0.098*	-0.098*	-0.098*	-0.098*	-0.098*
	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)
$NegativePert$	0.078	0.080	0.090	0.086	-0.028	-0.028	-0.028	-0.028	-0.028
	(0.032)	(0.034)	(0.032)	(0.033)	(0.024)	(0.025)	(0.024)	(0.024)	(0.023)
$ReligiousDiv$		0.256***	0.201*	0.195*	0.121	0.121	0.121	0.121	0.121
		(0.035)	(0.042)	(0.042)	(0.032)	(0.033)	(0.031)	(0.030)	(0.032)
$EthnicDiv$			0.111	0.105	0.094	0.094	0.094	0.094	0.094
			(0.041)	(0.039)	(0.029)	(0.028)	(0.028)	(0.028)	(0.029)
$Gender$				0.086	0.005	0.005	0.005	0.005	0.005
				(0.065)	(0.062)	(0.059)	(0.062)	(0.058)	(0.061)
$SocialPert$					0.705***	0.705***	0.705***	0.705***	0.704***
					(0.022)	(0.022)	(0.021)	(0.022)	(0.023)
$Constant$	-0.027	-0.977	-1.173	-1.309	-1.495*	-1.495*	-1.495*	-1.495*	-1.493*
	(0.155)	(0.168)	(0.165)	(0.166)	(0.148)	(0.138)	(0.138)	(0.146)	(0.143)
Dep Var: $R1Tokens$									
g	0.045	0.045	0.045	0.045	0.045	0.051	0.046	0.039	-0.002
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
ToM	0.057	0.057	0.057	0.057	0.057	0.055	0.043	0.047	0.089
	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
$NegativePert$	-0.528***	-0.528***	-0.528***	-0.528***	-0.528***	-0.527***	-0.531***	-0.530***	-0.535***
	(0.029)	(0.030)	(0.030)	(0.030)	(0.028)	(0.030)	(0.031)	(0.030)	(0.031)
$ReligiousDiv$						0.029	0.062	0.054	0.031
						(0.036)	(0.043)	(0.045)	(0.044)
$EthnicDiv$							-0.068	-0.074	-0.073
							(0.039)	(0.039)	(0.038)
$Gender$								0.096	0.069
								(0.072)	(0.074)
$SocialPert$									0.198**
									(0.032)
$Constant$	3.640***	3.640***	3.640***	3.640***	3.640***	3.528***	3.632***	3.434***	3.208***
	(0.163)	(0.171)	(0.173)	(0.169)	(0.169)	(0.186)	(0.181)	(0.179)	(0.187)
N	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000	122.000
AIC	1099.116	1226.533	1346.978	1327.367	1400.472	1402.330	1403.775	1404.191	1399.839
RSE	0.000	0.000	0.000	0.000	0.097	0.122	0.156	0.209	0.000
CFI	1.000	1.000	1.000	1.000	0.969	0.964	0.961	0.965	1.000

Note: Bootstrapped Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level. RSE=Residual square error, AIC = Akaike Information Criterion, CFI = Comparative Fit Index.

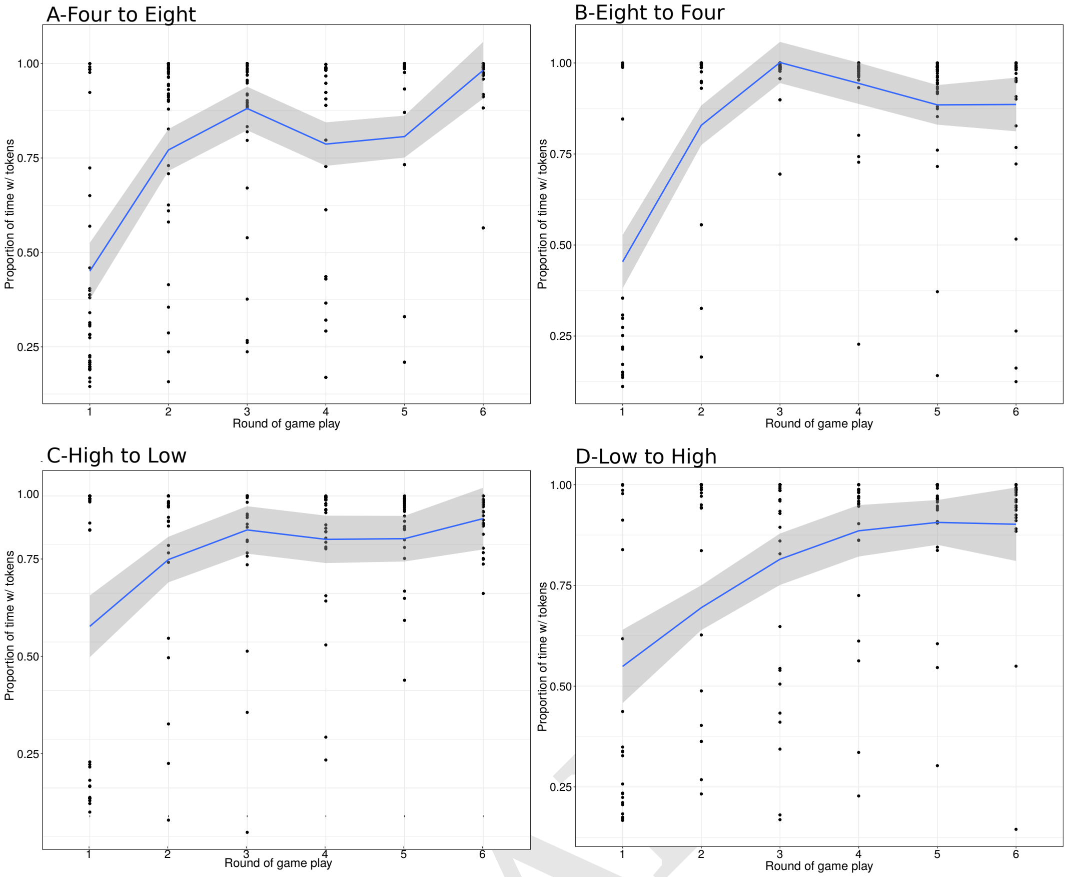


Fig. S3. Raw best fitting collective action learning curves.

4. Collective Action Learning Curves

To evaluate the effects of g and ToM on collective action learning curves in each treatment, we followed a three step analysis. First, we regressed $Round$ of game play on $Time$ separately for each treatment and compared the fit of a linear, quadratic and cubic model in each treatment. This allowed us to assess our predictions that negative perturbation treatments display cubic collective action learning curves, and that the positive perturbation treatments display quadratic collective action learning curves. Each regression was run using the general least squares method and fit by maximum likelihood estimation. Specifically, we compared three regressions:

$$Time = \beta_0 + \beta_1 Round + \epsilon \qquad [7]$$

$$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \epsilon \qquad [8]$$

$$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \beta_3 Round^3 + \epsilon \qquad [9]$$

where $Round$ is the round of game play in the six session time-series, β_0 is the intercept (mean $Time$ of Round 1), and $\beta 1 - 3$ describe how the amount of $Time$ with tokens on the screen changes as $Round$ of game play changes.

Tables S9–S12 compare the fits of the linear, quadratic and cubic models for each treatment. Fig. S3 displays the best fitting curve for each treatment. Consistent with expectations documented in the main paper, a cubic collective action learning curve best fits the negative perturbation treatments in which collective action becomes harder ('four to eight' and 'high to low' treatments). Conversely, in the positive perturbation treatments a cubic curve best fits the data in the 'eight to four' treatment. This pattern contradicts our expectation of a quadratic curve. In this instance, collective action becomes easier due to a decline in group size and groups perform a little worse, and then some groups improve again. In the 'low to high treatment', a quadratic curve best fits the data, and this is consistent with our expectations.

Second, after determining the fit of the best collective action learning curve, we ran regression models with interaction terms to assess how increases or decreases away from the mean g and ToM of groups affects collective action learning curves. Specifically, in the resource perturbation experiments ('high to low' and 'low to high'), we used general least squares regression fit by maximum likelihood to assess the effects of g and ToM on $Time$. In the case of the 'high to low' treatment, we fit the following model,

Table S9. A comparison of the fit of linear, quadratic and cubic collective action learning curves in the four to eight treatment

	call	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Linear	$Time = \beta_0 + \beta_1 Round + \epsilon$	1	3.00	38.85	49.14	-16.43			
Quad.	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \epsilon$	2	4.00	31.84	45.56	-11.92	1 vs 2	9.01	0.00
Cubic	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \beta_3 Round^3 + \epsilon$	3	5.00	8.74	25.89	0.63	2 vs 3	25.11	0.00

Table S10. A comparison of the fit of linear, quadratic and cubic collective action learning curves in the high to low treatment

	call	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Linear	$Time = \beta_0 + \beta_1 Round + \epsilon$	1	3.00	-25.89	-16.63	15.95			
Quad.	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \epsilon$	2	4.00	-29.61	-17.26	18.80	1 vs 2	5.71	0.02
Cubic	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \beta_3 Round^3 + \epsilon$	3	5.00	-31.76	-16.32	20.88	2 vs 3	4.15	0.04

$$Time = \beta_0 + \beta_1 Round - \beta_2 Round^2 + \beta_3 Round^3 + \beta_4 g + \beta_5 ToM + \beta_6 Round * g * ToM + \beta_7 Round^2 * g * ToM + \beta_8 Round^3 * g * ToM + \epsilon. \quad [10]$$

Similarly, for the 'low to high' treatment we fit,

$$Time = \beta_0 + \beta_1 Round - \beta_2 Round^2 + \beta_3 g + \beta_4 ToM + \beta_5 Round * g * ToM + \beta_6 Round^2 * g * ToM + \epsilon. \quad [11]$$

In both cases, we used the effects package in R to calculate the marginal effects of the interactions of $Round$, g, and ToM. Tables S13 and S14 display the regression coefficients, standard errors and significance of coefficients from equations (8) and (9). In addition to the main models described in these equations, we also ran robustness checks, adding and deleting the same control variables of religious diversity, ethnic diversity and the percent of a group composed of females noted in Section #3 above.

Table S13 illustrates that the results for the 'high to low' treatment discussed in the main text are robust to adding and deleting control variables. In this 'negative treatment' in which a decline in growth rate during round four should make collective action to sustain the resource harder, we observe increasing, decreasing and then increasing marginal returns consistent with our expectation. However, when adding the intelligence variables to the regression models, only the linear term for $Round$ consistently displays a significant effect. Consistent with the FIP, g and ToM both have positive direct effects on $Time$. These effects are observed across all regression models and are consistent with the results presented by Baggio and colleagues (2). However, while the effect of ToM is significant across all regression models, the significance of g does increase above p=0.1 in four of eight regression models. Finally, the interaction of g, ToM, and $Round$ is negative and significant across all regression models as discussed in the main text of the paper.

Table S14 illustrates that the results for the 'low to high' treatment discussed in the main text are robust to adding and deleting control variables. In this 'positive treatment' in which a decline in growth rate during round four should make collective action to sustain the resource easier, we observe increasing and then decreasing marginal returns consistent with our expectation. However, when adding the intelligence variables to the regression models, the $Round$ variables do not consistently display a significant effect. Partly consistent with the FIP, g has a positive direct effect on $Time$. This effect is observed across all regression models and is consistent with the results presented by Baggio and colleagues (2). However, the effect is not statistically significant across all regression models. The effect of ToM is near zero in all regression models and never significant.

In the group size experiments we ran mixed effects regression models fit by maximum likelihood estimation. These models partially replicate equations (8) and (9); however, we allowed the intercepts and the slopes of the polynomials for the $Round$ variable to vary randomly by groups of eight. In these treatments we ran mixed effects models to control for the effect that groups of eight may have on groups of four. As noted, we conducted our analysis on groups of four in the group size experiments as we tracked groups of four through all six rounds. Specifically, we compared the fits of six regression models: An intercept model; random intercept model for groups of eight; TimeRI which models the effects of $Round$ on $Time$ with random intercepts for groups of eight; TimeRI-Intell1 replicates TimeRI above with random intercepts for groups of eight and adds g, ToM and $Round$*G*ToM; TimeRI-Intell2 replicates TimeRI-Intell1 but includes an autocorrelation error function; and TimeRS replicates TimeRI-Intell2 with the addition of random slopes for the round polynomial for groups of eight. Tables S15 & S16 present the results of an ANOVA comparing these six mixed effects models in each respective treatment. In both treatments, the TimeRS model was the best fit, and we calculated the marginal effect of $Round$, g and ToM presented in the main body of the paper from these models.

Table S11. A comparison of the fit of linear, quadratic and cubic collective action learning curves in the eight to four treatment

	call	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Linear	$Time = \beta_0 + \beta_1 Round + \epsilon$	1	3.00	39.40	49.35	-16.70			
Quad.	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \epsilon$	2	4.00	-11.72	1.55	9.86	1 vs 2	53.11	0.00
Cubic	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \beta_3 Round^3 + \epsilon$	3	5.00	-25.42	-8.83	17.71	2 vs 3	15.70	0.00

Table S12. A comparison of the fit of linear, quadratic and cubic collective action learning curves in the low to high treatment

	call	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Linear	$Time = \beta_0 + \beta_1 Round + \epsilon$	1	3.00	26.97	36.00	-10.49			
Quad.	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \epsilon$	2	4.00	22.30	34.35	-7.15	1 vs 2	6.67	0.01
Cubic	$Time = \beta_0 + \beta_1 Round + \beta_2 Round^2 + \beta_3 Round^3 + \epsilon$	3	5.00	21.99	37.04	-5.99	2 vs 3	2.32	0.13

Table S13. General least squares regression of the collective action learning curve in the ‘high to low’ treatment

Model Coefficients

Predictors	Intell. Model	Control1	Control2	Control3	Control4	Control5	Control6	Control7
g	0.01**	0.01**	0.01**	0.005	0.01*	0.006	0.007	0.007
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
ToM	0.031***	0.036***	0.034***	0.038***	0.036***	0.037***	0.037***	0.037***
	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)
$Round$	2.93***	2.93***	2.93***	2.93***	2.93***	2.93***	2.93***	2.93***
	(0.507)	(0.508)	(0.509)	(0.503)	(0.501)	(0.504)	(0.499)	(0.500)
$Round^2$	-0.68	-0.68	-0.68	-0.68	-0.68	-0.68	-0.68	-0.68
	(0.507)	(0.508)	(0.509)	(0.503)	(0.501)	(0.504)	(0.499)	(0.500)
$Round^3$	0.63	0.63	0.63	0.63	0.63	-1.505***	0.63	0.63
	(0.507)	(0.508)	(0.509)	(0.503)	(0.501)	(0.504)	(0.499)	(0.500)
$Round : g : ToM$	-0.012***	-0.012***	-0.012***	-0.012***	-0.012***	-0.012***	-0.012***	-0.012***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$Round^2 : g : ToM$	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$Round^3 : g : ToM$	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$Gender$		0.04			0.05	0.05		-0.02
		(0.06)			(0.07)	(0.06)		(0.07)
$ReligousDiv$			0.008		-0.003		0.08*	0.09*
			(0.03)		(0.03)		(0.04)	(0.05)
$EthnicDiv$				-0.06*		-0.06*	-0.12***	-0.12**
				(0.03)		(0.03)	(0.04)	(0.04)
Constant	0.36***	0.33***	0.34***	0.48***	0.33***	0.45***	0.43***	0.43***

Model Fit

N	162	162	162	162	162	162	150	
AIC	-82.1	-80.72	-80.18	-83.78	-78.70	-82.54	-85.75345	-83.83
RSE	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17

Note: Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level. RSE=Residual square error.

Table S14. General least squares regression of the collective action learning curve in the ‘low to high’ treatment

Model Coefficients

Predictors	Intell. Model	Control1	Control2	Control3	Control4	Control5	Control6	Control7
g	0.018**	0.017**	0.016*	0.014	0.015*	0.014	0.014	0.013
	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)
ToM	-0.006	-0.005	-0.004	-0.006	-0.003	-0.005	-0.005	-0.003
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.01)
$Round$	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
	(0.72)	(0.72)	(0.72)	(0.72)	(0.72)	(0.72)	(0.72)	(0.73)
$Round^2$	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42
	(0.72)	(0.72)	(0.72)	(0.72)	(0.72)	(0.72)	(0.72)	(0.73)
$Round : g : ToM$	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$Round^2 : g : ToM$	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$Gender$		-0.08			-0.08	-0.07		-0.07
		(0.12)			(0.11)	(0.11)		(0.11)
$ReligousDiv$			-0.049		-0.049		-0.029	-0.03
			(0.057)		(0.057)		(0.068)	(0.068)
$EthnicDiv$				-0.05		-0.04	-0.034	-0.025
				(0.05)		(0.05)	(0.065)	(0.066)
Constant	0.42**	0.49**	0.50**	0.54**	0.57**	0.59***	0.55**	0.60**

Model Fit

N	150	150	150	150	150	150	150	
AIC	22.10	23.46	23.30	23.20	24.66	24.78	24.99	26.50
RSE	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24

Note: Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level.

Table S15. ‘Four to eight’ multilevel model ANOVA table

	call	df	AIC	BIC	logLik	Test	L.Ratio	p-value
intercept	gls(model = Time ˜ 1	2.00	87.97	94.83	-41.98			
randomIntercept	lme.formula(fixed = Time ˜ 1, random = ˜1 | Gp8ID)	3.00	74.87	85.16	-34.44	1 vs 2	15.10	0.00
TimeRI	lme.formula(fixed = Time ˜ poly(Round, 3), random = ˜1 | Gp8ID)	6.00	-25.64	-5.07	18.82	2 vs 3	106.52	0.00
TimeRI_Intell1	lme.formula(Time ˜ poly(Round, 3) + g + ToM + poly(Round, 3):g:ToM, random = ˜1 | Gp8ID")	11.00	-42.53	-4.80	32.26	3 vs 4	26.88	0.00
TimeRI_Intell2	lme.formula(fixed = Time ˜ poly(Round, 3) + g +ToM + poly(Round, 3):g:ToM, random = ˜1 | Gp8ID, correlation = corAR1())	12.00	-59.97	-18.82	41.98	4 vs 5	19.44	0.00
TimeRS	lme.formula(fixed = Time ˜ poly(Round, 3) + g + ToM + poly(Round, 3):g:ToM, random = ˜poly(Round, 3) | Gp8ID, correlation = corAR1())	21.00	-57.52	14.50	49.76	5 vs 6	15.55	0.08

Table S16. ‘Eight to four’ multilevel ANOVA

	call	df	AIC	BIC	logLik	Test	L.Ratio	p-value
intercept	gls(model = Time ˜ 1)	2.00	70.79	77.43	-33.40			
randomIntercept	lme.formula(fixed = Time˜ 1 random = ˜1 |Gp8ID)	3.00	71.84	81.79	-32.92	1 vs 2	0.95	0.33
TimeRI	lme.formula(fixed = Time ˜ poly(Round, 3), random = ˜1 | Gp8ID)	6.00	-31.97	-12.06	21.98	2 vs 3	109.81	0.00
TimeRI_Intell1	lme.formula(fixed = Time ˜ poly(Round, 3) + g + ToM + poly(Round, 3):g:ToM, random = ˜1 | Gp8ID)	11.00	-29.70	6.80	25.85	3 vs 4	7.73	0.17
TimeRI_Intell2	lme.formula(fixed = Time ˜ poly(Round, 3) + g + ToM + poly(Round, 3):g:ToM, random = ˜1 | Gp8ID, correlation = corAR1())	12.00	-33.88	5.94	28.94	4 vs 5	6.18	0.01
TimeRS	lme.formula(fixed = Time ˜ poly(Round, 3) + g + ToM + poly(Round, 3):g:ToM, random = ˜poly(Round1, 3) | Gp8ID, correlation = corAR1())	21.00	-85.14	-15.46	63.57	5 vs 6	69.26	0.00

Table S17 illustrates the fixed, random effects and model fits of the 'four to eight' mixed effects regression for eight regression models. The first model, as above, is the Intelligence Model (TimeRS in Table S15). The remaining seven control models add and delete gender composition, religious diversity and ethnic diversity from the TimeRS mixed effects regression model specified in Table S15. Table S17 illustrates that the fixed effects of g and ToM are positive across all regression models. Further, the interaction between $Round$, g and ToM is negative and significant across all regression models. The interaction between $Round^2$, g and ToM is also negative and significant across all regression models presented in Table S17.

Table S17. The fixed and random effects coefficients of the collective action learning curve in the 'four to eight' treatment

Fixed Coefficients

Predictors	Intell. Model	Control1	Control2	Control3	Control4	Control5	Control6	Control7
g	0.024***	0.024***	0.026***	0.020***	0.026***	0.024***	0.025***	0.026***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.009)
ToM	0.012*	0.013*	0.01	0.013*	0.009	0.013*	0.010	0.010
	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)	(0.007)	(0.007)	(0.008)
$Round$	4.15***	4.15***	4.13***	4.15***	4.13***	4.14***	4.13***	4.13***
	(0.592)	(0.594)	(0.592)	(0.503)	(0.519)	(0.593)	(0.590)	(0.592)
$Round^2$	-1.95***	-1.95***	-2.00***	-1.96***	-2.00***	-1.96***	-2.00***	-2.00***
	(0.585)	(0.586)	(0.584)	(0.582)	(0.585)	(0.587)	(0.584)	(0.585)
$Round^3$	1.49**	1.49**	1.44**	1.48**	1.44**	1.48*	1.44**	1.44**
	(0.628)	(0.629)	(0.628)	(0.628)	(0.630)	(0.630)	(0.629)	(0.630)
$Round : g : ToM$	-0.015***	-0.015***	-0.015***	-0.015***	-0.015***	-0.015***	-0.015***	-0.015***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$Round^2 : g : ToM$	0.007**	0.007**	0.008**	0.008**	0.008**	0.008**	0.008**	0.008**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$Round^3 : g : ToM$	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$Gender$		0.011			0.011	0.011		-0.011
		(0.066)			(0.066)	(0.062)		(0.066)
$ReligousDiv$			0.04		0.045		0.041	0.044
			(0.04)		(0.047)		(0.045)	(0.048)
$EthnicDiv$				0.02		0.022	0.021	0.021
				(0.03)		(0.031)	(0.031)	(0.032)
Constant	0.12	0.12	0.07	0.07	0.07	0.11	0.06	0.06
Random Effects								
$\theta^2_{Intercept}$	0.12	0.12	0.13	0.12	0.12	0.13	0.13	0.13
θ^2_{Round}	0.66	0.67	0.70	0.66	0.7	0.66	0.69	0.70
$\theta^2_{Round^2}$	0.63	0.63	0.66	0.64	0.66	0.64	0.67	0.66
$\theta^2_{Round^3}$	0.95	0.95	0.99	0.95	0.99	0.95	0.99	0.99
Model Fit								
N	228	228	228	228	228	228	228	228
AIC	-57.51	-55.52	0.13	-56.01	5.66	-54.04	6.77	12.29
Log Likelihood	49.75	49.76	21.93	50.00	20.16	50.02	19.61	17.85

Note: Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level.

Fig. S4 illustrates the predicted fits of the Intelligence regression model with random intercepts and slopes overlaying the actual data for every group of four and eight. There are 38 groups of four in the dataset and 19 groups of eight. Each panel of the graphic describes how $Time$ changed over six rounds for each group of four, and each pair of windows illustrates groups of eight. For example, during rounds 1-3, groups 1 and 2 (panels 1 and 2 in the graphic) played the game as a group of four. In round four, these two groups combined into a group of eight. This pattern holds for the entire graphic, groups 3 and 4 became a group of 8 in round 4, 5 and 6 the same, and so on. The bottom line is that a cubic function is a good fit for many groups of four, with a decline in performance following the shift to a group of eight, and then an increase in performance as the group of eight readjusted to the larger group size.

Table S18 illustrates the fixed, random effects and model fits of the 'eight to four' mixed effects regression for eight regression models. The first model, as above, is the Intelligence Model (TimeRS in Table S15). The remaining seven control models

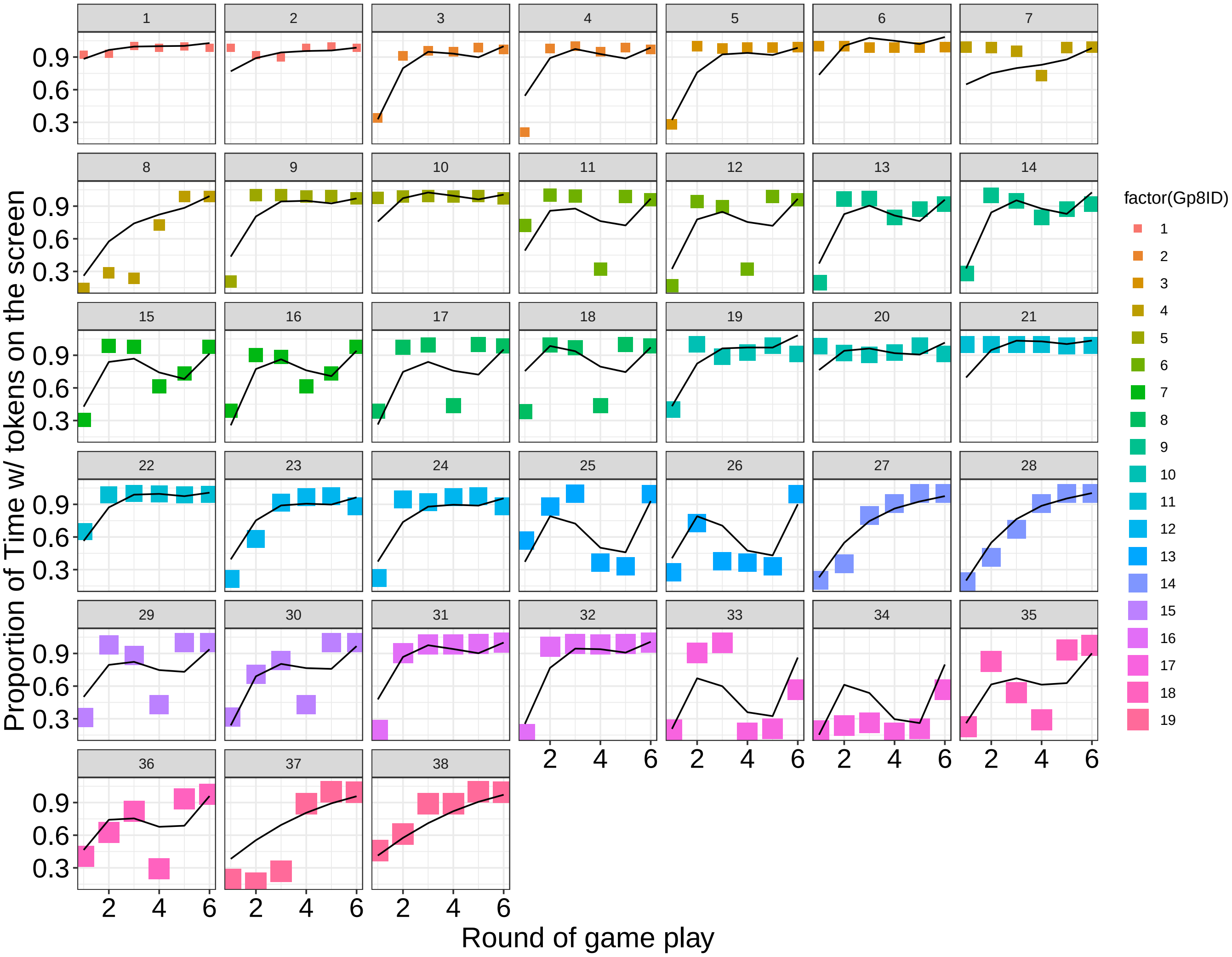


Fig. S4. The predicted curves of the mixed effects model in the four to eight treatment plotted over the actual data.

add and delete gender composition, religious diversity and ethnic diversity from the TimeRS mixed effects regression model specified in Table S15. Table S18 illustrates that the fixed effects of g is positive and significant across all regression models. Further, the effects of the *Round* polynomial are statistically significant across all regression models. The interaction between the *Round* terms, g and ToM is not statistically significant in any of the models. As with the 'low to high' treatment reported above, ToM has a negligible fixed effect across regression models reported in S18.

Fig. S5 illustrates the predicted fits of the Intelligence regression model with random intercepts and slopes overlaying the actual data for every group of four and eight. There are 34 groups of four in the dataset and 17 groups of eight. Each panel of the graphic describes how *Time* changed over six rounds for each group of four, and each pair of windows illustrates groups of eight. For example, during rounds 1-3, groups 1 and 2 (panels 1 and 2 in the graphic) played the game as a group of eight. In round four, these two groups split into two groups of four. This pattern holds for the entire graphic; groups 3 and 4 became a group of 8 in round 4, groups 5 and 6 the same, and so on. The bottom line is that groups display a lot of variability after the group size change from eight to four. Over the fist three rounds, almost every group of eight improves their collective action to sustain the resource without fail. However, once the positive perturbation hits and group size decreases to four, some groups continue to sustain the resource well, while others perform more poorly. This visual observation is consistent with the results presented in the path analysis in which treatment type (0=negative, 1=positive) has a negative and significant effect on both the robustness of *Time* and *Tokens*. Groups in the positive treatments display more variability round to round and between groups than groups in the negative perturbation treatments.

Table S18. The fixed and random effects coefficients of the collective action learning curve in the ‘eight to four’ treatment

Fixed Coefficients

Predictors	Intell. Model	Control1	Control2	Control3	Control4	Control5	Control6	Control7
g	0.016***	0.015**	0.017***	0.016***	0.016***	0.015**	0.017***	0.016**
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
ToM	-0.0007	0.00008	-0.0007	-0.0001	0.009	-0.002	-0.002	-0.001
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$Round$	1.22**	1.22**	1.22**	1.22**	1.22**	1.22**	1.21**	1.21**
	(0.583)	(0.584)	(0.584)	(0.583)	(0.585)	(0.584)	(0.584)	(0.585)
$Round^2$	-2.04***	-2.05***	-2.04***	-2.03***	-2.04***	-2.03***	-2.03***	-2.04***
	(0.508)	(0.509)	(0.510)	(0.509)	(0.511)	(0.510)	(0.510)	(0.511)
$Round^3$	0.73*	0.73*	0.73*	0.74*	0.73*	0.74*	0.74*	0.74*
	(0.44)	(0.44)	(0.44)	(0.44)	(0.44)	(0.44)	(0.44)	(0.44)
$Round : g : ToM$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$Round^2 : g : ToM$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$Round^3 : g : ToM$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$Gender$		0.049			0.034	0.052		0.038
		(0.054)			(0.066)	(0.054)		(0.054)
$ReligousDiv$			-0.051*		-0.047		-0.047	-0.042
			(0.03)		(0.031)		(0.031)	(0.032)
$EthnicDiv$				-0.023		-0.025	-0.017	-0.019
				(0.028)		(0.027)	(0.020)	(0.027)
Constant	0.44***	0.44***	0.47***	0.46***	0.46***	0.46***	0.48***	0.48***

Random Effects

$\theta^2_{Intercept}$	0.07	0.07	0.067	0.07	0.067	0.07	0.068	0.067
θ^2_{Round}	1.53	1.53	1.53	1.53	1.53	1.53	1.53	1.53
$\theta^2_{Round^2}$	1.12	1.12	1.12	1.12	1.12	1.12	1.12	1.12
$\theta^2_{Round^3}$	0.56	0.56	0.55	0.56	0.55	0.56	0.55	0.55

Model Fit

N	204	204	204	204	204	204	204	204
AIC	-85.02	-83.85	-85.57	-83.76	-83.97	-82.75	-83.96	-82.48
Log Likelihood	63.51	63.92	64.78	63.88	64.98	64.37	64.98	65.24

Note: Standard errors reported in parenthesis * = significant at the 90% level, ** = significant at the 95% level, and *** = significant at the 99% level.

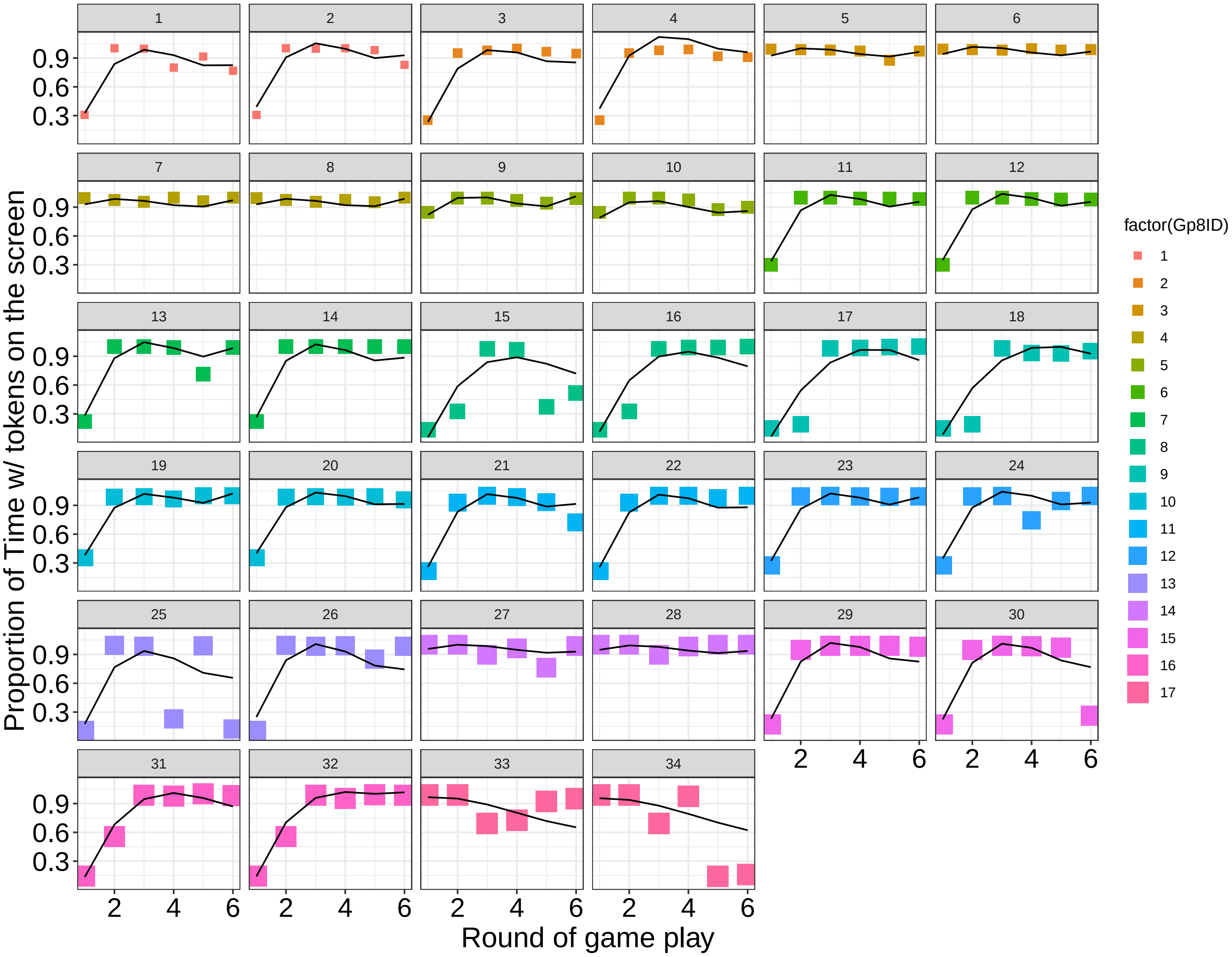


Fig. S5. The predicted curves of the mixed effects model in the eight to four treatment plotted over the actual data.

5. Recruitment for the Experiments

We recruited 550 undergraduate students from two universities: Utah State University (USU) and the University of Texas at San Antonio (UTSA). 216 participants participated in the resource treatments, and 334 individuals participated in the group size treatments. We recruited students at UTSA from the introductory psychology participation pool. The introductory psychology participation pool draws students from approximately 8 introductory psychology courses at over 100 students per course where students complete studies for partial credit toward course completion. These introductory courses draw students from across the university. At USU, we recruited students from introductory sociology and anthropology courses that have over a 150 students enrolled per course. These courses are general requirement courses and draw students from all majors represented at the university. Participation in the study was voluntary and students could withdraw participation at any time. All participants received course credit for participation and a cash payout that depended on how many tokens they collected over six rounds.

A. External Validity. Experimental studies, especially when performed within universities, can have issues with external validity: How well does the sample population represent the population of interest? Here we compare our sample population with the U.S. population to assess how well we are capturing wider characteristics of interest. We use the CIA Factbook, retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/us.html to compare general demographic data, such as ethnic diversity, religious diversity, gender composition and age.

A.1. Demographics. The tables below portray our sample summary statistics as well as distribution for Ethnic and Religious Diversity. According to the data retrieved from the CIA factbook, within the U.S. Ethnic Diversity is, overall, 0.970, however, including Hispanic, ethnic diversity displays a value of 1.289. Our sample population displays an ethnic diversity index of 0.55 on average, while 50% of the groups within our sample population display an ethnic diversity index of 0.562. The sample population is thus more homogeneous, from an ethnic point of view, than the U.S. population. However, it is also true that over 30% of groups participating within the experiment display an ethnic diversity in line with the U.S. population (26.61% of groups had an ethnic diversity index of 1.04 and 6.45% of groups had an ethnic diversity index of 1.386, both in line with the U.S. data between 0.97 and 1.23 depending on whether Hispanic is included or not). Note that the CIA factbook does not directly report Hispanic population but gives estimates, here the estimates were used to calculate Hispanic population by proportionally reducing black and white populations.

Summary Statistic	Ethnic Diversity	Religious Diversity
N	122	122
Mean	0.550	0.867
Std. Dev.	0.042	0.038
25th Percentile	0.000	0.562
Median	0.562	1.040
75th Percentile	1.040	1.040
Min	0.000	0.000
Max	1.386	1.386

Ethnic Diversity	Percent
0.000	36.29
0.562	21.77
0.693	8.87
1.040	26.61
1.386	6.45

On average, we find that our religious diversity is very much in line with the average religious diversity within the wider U.S. population. In fact, our average religious diversity metric (0.867, median 1.04) is similar with the average religious diversity index calculated for the whole U.S., where the religious diversity index is 0.79 grouping Christians together, and 1.38 keeping Christian religious affiliations separated. Note that these results are very much in line with our sample, where over 33.06% of groups displayed a religious diversity index of 1.04, and 24.19% of the groups displayed a diversity index of 1.386.

Religious Diversity	Percent
0.000	10.48
0.652	27.42
0.693	4.84
1.040	33.06
1.386	24.19

With respect to Age, our sample is skewed towards a younger population. This is a bias that is almost impossible to eliminate when performing behavioral experiments within the university context, where undergraduate populations are sampled. However, the results of the study are still valuable. The results provide a firmer basis for conducting more experiments with a

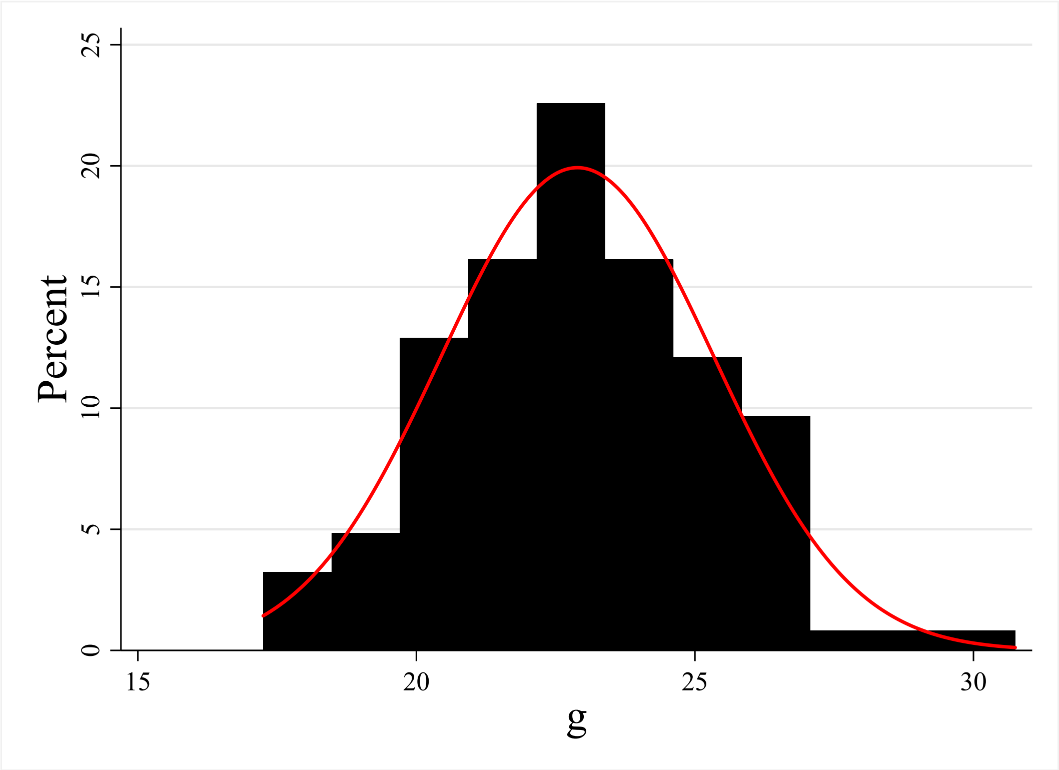


Fig. S6. Distribution of group general intelligence ($avg\ g$) within the sample population.

wider range of age cohorts. Finally, our sample displays a higher gender imbalance as 62.5% of participants self-identified as females, compared to the general U.S. female population that represent 50.1% of the overall population.

Overall, although with some issues with respect to age and gender, our sample population reflects the diversity seen within the wider U.S. population.

A.2. g and ToM. With respect to g, as Fig. S6 shows, the participants in the experiment have on average higher g than the general population. With respect to ToM, this is difficult to assess as no US or other whole population data are available. However, ToM is moderately correlated with the personality trait of agreeableness (see (1) as well as Nettle and colleagues (9)). ToM in our sample population is depicted in Fig. S7.

6. Short Story Test

In the Short Story Task (SST), participants read "The End of Something", a short story by Ernest Hemingway, which presents a nuanced interaction between a romantic couple in which the male protagonist, Nick, starts an argument and breaks up with his girlfriend, Marjorie. Through the course of the story, the characters display sarcasm, non-verbal and indirect communication, higher-order emotions like guilt, and attempts to hide their intentions and feelings from one another.

According to Dodell-Feder (8), the goal of the SST was to

> "to design a new ToM task (the Short Story Task -SST-) that improved upon the limitations of existing ToM measures. More specifically, we aimed to create a task that (a) was sensitive to individual differences in ToM ability and did not suffer from ceiling effects, (b) incorporated a range of mental states of differing complexity, including epistemic states, affective states, and intentions to be inferred from a first- and second-order level, (c) used ToM stimuli representative of real-world social interactions, (d) required participants to utilize social context when making mental state inferences, (e) exhibited adequate psychometric properties, and (f) was quick and easy to administer and score." (8, p. 2)

A. Short Story Test instruction and questions. **Instructions to Participant**

Now you are going to read a short story called The End of Something. The story is only a few pages, but take your time reading it. Try to get a sense of what happens and what the relationships are between the characters. After you're finished, some questions will appear on the screen and you will be asked to answer them.

After story is read

1. Have you read this story before? [yes | no]

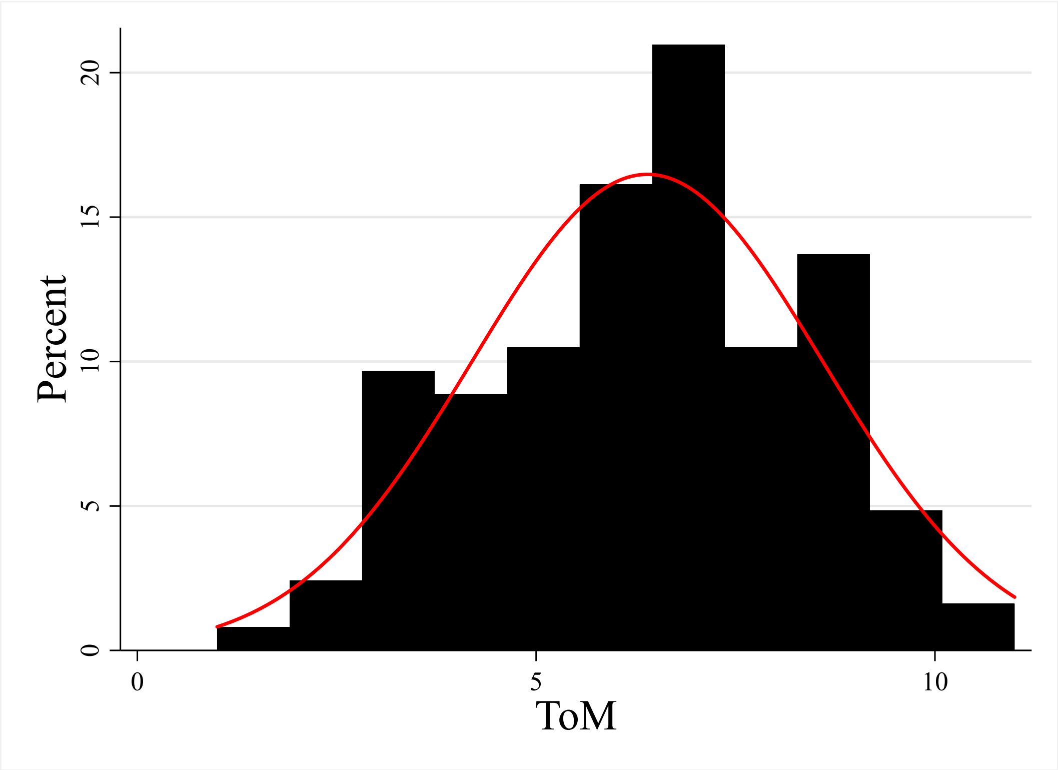


Fig. S7. Distribution of group ToM ($min\ ToM$) within the sample population

- IF YES
  - How long ago did you read it?
  - How well do you remember the story?
  - Did you read it for school or pleasure?
    - IF SCHOOL
      - What grade were you in?
      - What class was it for?

2. Is the story familiar to you? [yes | no]

- IF YES
  - Do you know anything about the story? What do you know about it?
  - Have you discussed the story with anyone?

Instructions to Participant Now I'm going to ask you some questions about the story. Here is a copy of the questions I'll be asking so you can read along. For most of the questions, there are no right or wrong answers and the questions can be answered with short responses. We're also interested in the character's thoughts, feelings and intentions when it applies to the question.

Questions

1. In just a few sentences, how would you summarize the story
2. What do Nick and Marjorie observe on the shoreline as they are rowing to the point to set their fishing lines?
3. What does Nick mean when he says, "They aren't striking?"
4. Nick and Marjorie have a pail of perch for what purpose?
5. Do Marjorie's actions suggest that she is experienced or inexperienced at fishing? What makes you say that?
6. Why does Nick say to Marjorie, "You know everything"?
7. Why does Marjorie reply, "Oh Nick, please cut it out! Please, please don't be that way!"?

8. Why is Nick afraid to look at Marjorie?

9. What does Nick mean when he says, "It isn't fun anymore"?

10. Why does Marjorie sit with her back toward Nick when she asks, "Isn't love any fun?"?

11. Why does Marjorie take the boat and leave and what is she feeling at that moment?

12. Who is Bill and what does he reveal when he asks Nick, "Did she go alright? ... Have a scene?"?

13. What is Nick feeling when he says, "Oh, go away, Bill! Go away for a while"?

14. The story is called "The End of Something." What is the title referring to?

A.1. Scoring the Short Story Test. Scoring for the SST followed the methodology of (8). Three different coders coded the answer to the SST independently. We calculated Krippendorff's alpha using ordinal data using ReCal online. Krippendorff's alpha (ordinal) was 0.833, demonstrating a high level of coders agreement. As the SST was consistently coded with three coders, all questions that did not have 100% agreement among coders was coded as the score issued by the majority of coders. If a question did not have a majority (all coders issued different scores) disagreements were resolved via discussion between all coders.

The following is the coding sheet used by the coders. Explicit mental state reasoning (in bold) is the metric used to assess Social Intelligence.

- Comprehension: Sum scores of 5 comprehension questions (questions 2, 3, 4, 5, and 14). Ranges from 0 to 10.

- **Explicit mental state reasoning**: Sum scores of 8 mental state reasoning questions (questions 6, 7, 8, 9, 10, 11, 12, and 13). Ranges from 0 to 16.

- Spontaneous mental state inference: 1 score for spontaneous mental state question (question 1). Ranges from 0 to 1.

Following are examples of coding used to evaluate and score the three components of the test described above.

- Question 1: 1 = any mental state inference, even if it is wrong

- Question 2: 2 = any adjective + mill 1 = only mill 0 = anything else

- Question 5: 2 = experienced 2 or 1 = somewhat experienced / somewhat inexperienced 2 for good justification 1 for bad / no justification 1 or 0 = inexperienced 1 for good justification 0 for bad / no justification

- Question 6: 0 = anything that does not understand that he's being sarcastic, anything that thinks he's joking, anything that thinks that she does actually know everything

- Question 7: 2 = if they understood that he was giving her a hard time or doing something that was not intended to make her happy

- Question 8: 2 = anything that references her reaction / emotions 1 = anything that references his reaction / emotions without referencing hers 0 = no mention of an emotion

- Question 10: 2 = knows about break up / something bad (may include emotion) 1 = emotion with no knowledge of breakup / something bad 0 = No emotion, No knowledge about break up / something bad

- Question 11: 2 = (either the relationship is over or wanting space) AND negative emotion 1= upset OR wants space

- Question 12: 2 = Bill's relationship with Nick AND anything that references Bill's advanced knowledge 1 = Bill's relationship with Nick OR directly states Bill knew Nick was going to break up with Marjorie (Bill is not in the clearing while Nick and Marjorie fight and/or break up. He enters later.)

- Question 13: 2 = negative emotion referencing break up AND needs space / doesn't want to talk 1 = negative emotion 0 = no negative emotion, only wants space

- Miscellaneous: As long as a correct answer is present (even if a patently wrong answer is also present), give the score for the correct answer. Anything that is obviously wrong (outside of question 1) should be scored as 0.

7. Survey

Participants filled out the following exit survey to collect data researchers have identified as relevant for explaining performance on common pool resource experiments in the past (3, 18, 19). They survey also was used to collect basic demographic information.

1. Please report your age in years.
2. Please provide your current GPA
3. Please describe your religious affiliation, if any. Please be as specific as possible.
4. What is your primary language?
5. Please specify how you identify your race or ethnicity.
6. Please indicate the number of individuals who you call close friends? Please exclude family members and mere acquaintances.
7. How many individuals are in your total social network (i.e., close friends plus family members plus acquaintances)?
8. Please write the typical number of individuals who lived in your home while you were between the ages of 5-17?
9. Please estimate the median household income of the family in which you lived between ages of 5-17?
10. What is your college major or intended college major?
11. Please circle the descriptor that best describes your biological sex?
    - M
    - F
12. Did you understand the instructions of the exercises?
    - I did not understand anything I understood only a bit of the instructions
    - I understood half of the instructions
    - I understood most of the instructions
    - I understood everything
13. Do you think most people would try to take advantage of you if they had a chance, or would they try to be fair?
    - Would take advantage of you
    - Depends on situation
    - Would try to be fair
14. Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?
    - Try to be helpful
    - Depends on situation
    - Mostly just looking out for themselves
15. Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?
    - Most people can be trusted
    - Depends on situation
    - Can't be too careful in dealing with people
16. In the past year, did you do any volunteer activity through organizations; i.e. donate your time and energy not for pay?
    - Yes
    - No
17. Global warming is a fact and is mostly caused by emissions from vehicles and industrial facilities?
    - I completely agree

- I somewhat agree
- I have no opinion
- I somewhat disagree
- I completely disagree

18. Tell me whether the first statement or the second statement comes closer to your own views ? even if neither is exactly right.
    - Most people who want to get ahead can make it if they're willing to work hard.
    - Hard work and determination are no guarantee of success for most people.
19. Tell me whether the first statement or the second statement comes closer to your own views ? even if neither is exactly right.
    - The government should do more to help needy Americans, even if it means going deeper into debt.
    - The government today can't afford to do much more to help the needy
20. Here are two statements people sometimes make when discussing the environment and economic growth. Which of them comes closer to your own point of view?
    - Protecting the environment should be given priority, even if it causes slower economic growth and some loss of jobs.
    - Economic growth and creating jobs should be the top priority, even if the environment suffers to some extent.
21. What is the highest educational level that your Parents have attained?
22. What is your Father's occupation?
23. What is your Mother's occupation?

1. Freeman J, Coyle TR, Baggio JA (2016) The functional intelligences proposition. Personality and Individual Differences 99:46–55.
2. Baggio JA, et al. (2019) The importance of cognitive diversity for sustaining the commons. Nature communications 10(1):875.
3. Janssen MA, Holahan R, Lee A, Ostrom E (2010) Lab experiments for the study of social-ecological systems. Science 328(5978):613–617.
4. Frey MC, Detterman DK (2004) Scholastic assessment or g? the relationship between the scholastic assessment test and general cognitive ability. Psychological science 15(6):373–378.
5. Koenig KA, Frey MC, Detterman DK (2008) Act and general cognitive ability. Intelligence 36(2):153–160.
6. Coyle TR, Pillow DR (2008) Sat and act predict college gpa after removing g. Intelligence 36(6):719–729.
7. College board (2016) Concordance tables.
8. Dodell-Feder D, Lincoln SH, Coulson JP, Hooker CI (2013) Using fiction to assess mental state understanding: A new task for assessing theory of mind in adults. PLoS ONE 8(11):1–14.
9. Nettle D, Liddle B (2008) Agreeableness is related to social-cognitive, but not social-perceptual, theory of mind. European Journal of Personality 22(4):323–335.
10. Barrick MR, Stewart GL, Neubert MJ, Mount MK (1998) Relating member ability and personality to work-team processes and team effectiveness. Journal of applied psychology 83(3):377.
11. Meslec N, Aggarwal I, Curseu PL (2016) The Insensitive Ruins It All: Compositional and Compilational Influences of Social Sensitivity on Collective Intelligence in Groups. Frontiers in Psychology 7(MAY):1–7.
12. Ostrom E (1998) A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, 1997. American Political Science Review 92(01):1–22.
13. Baggio JA, Rollins ND, Pérez I, Janssen MA (2015) Irrigation experiments in the lab: Trust, environmental variability, and collective action. Ecology and Society 20(4):12.
14. Baggio JA, et al. (2016) Multiplex social ecological network analysis reveals how social changes affect community robustness more than resource depletion. Proceedings of the National Academy of Sciences 113(48):13708–13713.
15. Alesina A, Devleeschauwer A, Easterly W, Kurlat S, Wacziarg R (2003) Fractionalization. Journal of Economic growth 8(2):155–194.
16. Baggio JA, Papyrakis E (2010) Ethnic diversity, property rights, and natural resources. The Developing Economies 48(4):473–495.
17. Schill C, Lindahl T, Crépin AS (2015) Collective action and the risk of ecosystem regime shifts: Insights from a laboratory experiment. Ecology and Society 20(1).
18. Ostrom E (2005) Understanding Institutional Diversity. (Princeton University Press, Princeton).
19. Poteete A, Janssen M, Ostrom E (2010) Working together: collective action, the commons, and multiple methods in practice. (Princeton Univ Pr).