Conducting Large, Repeated, Multi-Game Economic Experiments using Mobile Platforms

Zhi Li,[1,2] Po-Hsuan Lin,[3,4] Si-Yuan Kong[4], Dongwu Wang[4], John Duffy[5*]

[1]Department of Public Finance, School of Economics,
the Wang Yanan Institute for Studies in Economics (WISE), Xiamen University,
422 Siming S Rd, Siming District, Xiamen, Fujian, China, 361005

[2] The MOE Key Laboratory of Econometrics, Xiamen University,
422 Siming S Rd, Siming District, Xiamen, Fujian, China, 361005

[3]Division of the Humanities and Social Sciences, California Institute of Technology,
1200 East California Blvd, MC 228-77, Pasadena, CA, USA, 91125

[4]MobLab Inc.,
380 N Halstead St, Pasadena, CA, USA, 91107

[5]Department of Economics, University of California, Irvine, CA, USA, 92697

[*]To whom correspondence should be addressed; E-mail: duffy@uci.edu.

One-sentence summary: We report on a breakthrough methodology for conducting large scale economic experiments, connecting laboratory with field research.

We demonstrate the possibility of conducting repeated, multi-game economic decision-making experiments with hundreds of subjects using entirely mobile platforms. Our experiment provides important proof-of-concept that such experiments are not only possible, but yield recognizable results as well as new insights, and are a promising way forward in the post-COVID-19 era, blurring the line between laboratory and field studies. Specifically, our findings from the 8 different experimental economics games and tasks replicate existing results despite the fact that subjects play those games/task in a specific order. We leverage our large subject population to study the effects of large ($n = 100$) **versus small** ($n = 10$) **group sizes on behavior. We can replicate several existing findings for small groups, but increases in group size are shown to matter for the robustness of those findings. Finally, we use our multi-game design to examine the consistency of strategic sophistication across games and for correlations in certain measures such as risk and social preferences between games.**

Introduction

Economic decision-making and game theory experiments have traditionally been conducted with small numbers of subjects in laboratory settings. However, recent advances in cloud-based platforms and mobile payments mean that laboratory-type studies can now be conducted outside of the traditional lab space and involve much larger numbers of participants. In this paper, we report on a break-through, incentivized experimental study involving more than 1,200 university students playing eight classic laboratory games or individual decision-making tasks. Our experiments took place during a summer camp at Xiamen University in July 2019 (Experiment 1, 633 subjects) and was repeated again with different Xiamen summer school participants in July

2020 (Experiment 2, 585 subjects) during the COVID-19 pandemic. In both sessions, subjects used their own smartphones, tablets, or laptops, and played games using the same cloud-based platform, MobLab (https://www.moblab.com/). They received payments based on their choices via Alipay, the mobile payment platform of Alibaba.

We have several aims in conducting and reporting on this experiment. First, we demonstrate proof of concept that such large scale experimentation can be done entirely using mobile platforms in-person or remotely and yield both recognizable results. Indeed our results with small groups of 2-10 subjects replicate findings from traditional laboratory studies despite the fact that subjects in our experiment play several different games in a sequence (a within-subjects design) as opposed to the between-subjects designs of most laboratory studies. We further demonstrate that our results are not substantially different if we conduct our large scale experiment *in-person*, as in the 2019 Experiment 1, or *remotely* as in the 2020 Experiment 2.[1] These encouraging findings support the prospect of conducting efficacious and incentivized human subject experiments remotely online during the COVID-19 pandemic and beyond without sacrificing comparability to results from traditional in-person laboratory experiments.

Second, we leverage our large subject pool to explore how group size differences of 10 subjects versus 100 subjects can matter for the play of three highly scalable classic laboratory games: the p-beauty contest game, a voter turnout game, and a linear public goods game. We find some significant differences in behavior between the large and small groups among the three games, suggesting that group size may be an important (and often overlooked) factor in evaluating laboratory findings.

Third, since participants also play five other 2-player games or individual decision tasks, we examine *correlations* in behavior across all eight games or tasks played to better understand the persistence of subject characteristics across those games. Specifically, in addition to the three games exploring group size effects, subjects also participated in a 2-player ultimatum bargaining game (*1*), an individual risk preference elicitation task similar to (*2*), a 2-player centipede game played 3 times (*3*), a 2-player trust game (*4*) and an individual 3-round real-effort task exploring gender differences in compensation schemes (*5*). Using this rich data set, we examine whether a measure of subject's strategic sophistication based on the first game they played, the p-beauty contest game explains their behavior in the other seven games/tasks, a novel cross-game analysis of of strategic sophistication. We find that the greater the strategic sophistication of a subject in the p-beauty contest game, the more likely that subject is to be a free rider in the public goods game or to offer very little as the first mover in the ultimatum game. More generally, we look for correlations in outcomes across all 8 games/tasks and find other interesting relationships such as a negative correlation between greater risk aversion and the incidence of free riding behavior.

Finally, we note that, to our knowledge, this study comprises two of the largest synchronous, repeated, multiple game/ decision-task experiments with paid human subjects that have ever been conducted. Our experiment is easily replicated by others as we use standard, pre-programmed games (that are easily configurable) available on the MobLab platform.

The idea that group size can matter for experimental results has not gone unnoticed. In one of the earliest studies of the public good game, (*6*) examined groups of size 4, 10, 40 and 100 and found that average contributions increased with the group size, e.g., from 10 to 100, but only if the marginal per capita return (MPCR) on contributions to the public good was low (0.30). (*7*) replicate this finding in comparisons of groups of size 4 and 8, but find that larger groups contribute less than smaller groups when the MPCR is high (0.75). (*8*) report on beauty contest games conducted among the readership of three newspapers resulting in sample sizes of 1,476 to 3,696 participants. (*9*) study voter turnout in the laboratory in small groups of size 3 as well as in larger groups of up to size 53, while (*10*) conduct a similar experiment involving groups of size 30 or 300 using an Mechanical Turk's online workforce. The findings from all of these studies is that group size can matter.[2] Still, these studies are typically conducted on different dates in time or on different populations (laboratory subjects versus newspaper readers, combining several laboratories at once) or lack other elements of control, e.g., newspaper readers can discuss their choices with one another. Thus, an important contribution of our paper is that we conduct the large versus small group treatments *simultaneously* via random assignment of members of a single population to either the small or large group sizes. Thus, our design does not mix populations making better use of random assignment and therefore resulting in a more controlled test of group size effects.

Regarding our other research objective of studying interaction effects *across* games there is also prior work by experimental economists. For instance, (*11*) identify correlations between risk aversion, time preferences, loss aversion and ambiguity aversion. (*12*) find non-trivial correlations between "econographic" variables such as risk aversion, over-confidence, etc. Some studies have also tried to identify correlations between observed strategic sophistication levels and the personality traits. Generally speaking, more sophisticated players are more likely to have greater academic training (*13*), higher cognitive ability (*14, 15*), and higher intelligence (*16*). By contrast, there are relatively few studies examining correlations in *strategic behavior* across games. For instance, (*17*) and (*18*) attempt to identify the behavioral correlations between guessing games and the dominance-solvable normal form games but they cannot find robust correlations. Most of the studies exploring correlations in strategic behavior across games involve only *two* games, whereas we look for interaction effects across eight different games or tasks that have been widely studied.

Results

In this section, we present results germane to our specific research aims following the order in which the games were played in our experiment. Methodological details are provided in the Methods section and in the Supplementary Material F.

Large versus Small Group Size Effects

The main treatment variable of our first 3 games, for which group size is scalable, is *Large* ($N = 100$) vs. *Small* ($N = 10$) group sizes. For each game or task, we relate our findings to those found in the literature.

Beauty Contest Game

In our first game, initially studied by (*19*), participants in each group of size N simultaneously and without communication guess a number in the interval [0,100]. The winning guess is the number closest, in absolute value, to p times the group average. We set $p = 2/3$ and the game was played for 3 repetitions (rounds) among unchanging members of a group of size 10 or 100. The winner in each group earned 20 points per round (the prize is evenly split for a tie); everyone else earned 0 points.

Regardless of group size, iterative elimination of dominated strategies results in the equilibrium prediction that all N players will guess 0.

Figure 1 shows the cumulative distributions of guesses across experiments, treatments, and rounds. We use Kolmogorov–Smirnov (KS) tests to show that guesses in large groups are distributed *lower* and thus *closer* to the equilibrium prediction of 0 than in small groups by round 2 of Experiment 1 ($KS = 0.2709$, $p < 0.001$) and by round 3 of Experiment 2 ($KS = 0.2450$, $p < 0.001$), although the difference is not statistically significant in round 1 (Experiment 1: $KS = 0.1457$, $p = 0.080$; Experiment 2: $KS = 0.0924$, $p = 0.251$). This finding is rather intuitive: in larger groups, subjects are primed by the greater competitive pressure to iterate their reasoning further than they would in smaller, less competitive groups.

Other studies (e.g. (*8*)) have found no statistically significant group size effects in this game, but they have typically compared different subject populations and/or one-shot games. The most similar study (*20*) compared guesses in a $p = .7$ beauty contest game by small ($N = 3$) versus

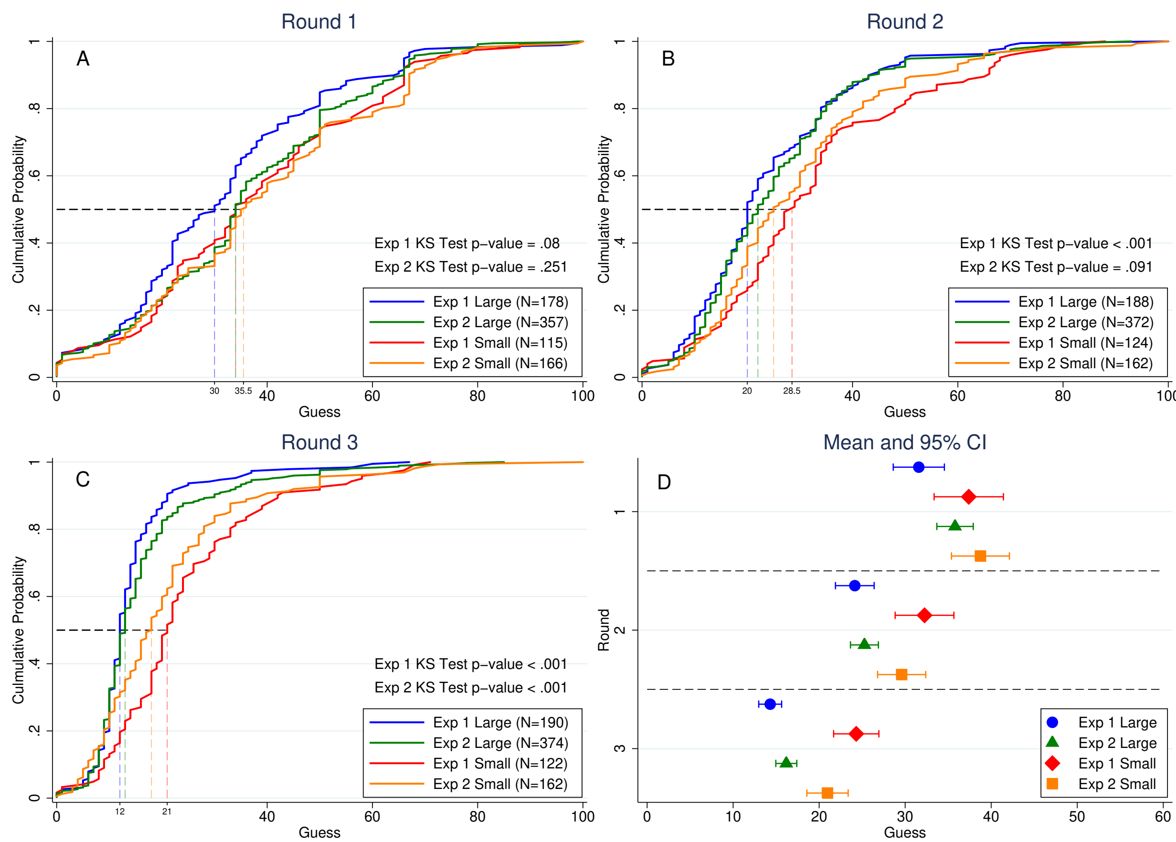

Fig. 1: Cumulative density function of guesses across different rounds. (A-C) The distribution of guesses in large groups and small groups from Experiment 1 and 2. The numbers of observations are shown in the legends and the median guesses of large and small groups are labeled in dashed lines. The p-values of the Kolmogorov–Smirnov tests for different groups are provided at the bottom of each figure. (D) The means and 95% CI for different groups from round 1 to round 3.

large ($N = 7$) groups over 10 rounds and found that larger groups played the equilibrium prediction more often than did smaller groups, but their result was not statistically significant.

Voter Turnout Game

Our second game is based on the experimental voter turnout study of (*9*) in which there are two teams of different sizes, in a ratio of approximately 2:1 membership. In Small groups, the majority team has 7 members while the minority team has 3 members. In Large groups, the majority has 67 members to the minority's 33 members. Members of each team have to simultaneously and without communication decide whether or not to vote. The team with the most votes wins a prize of 100 points, the losing team gets 0 points, and in the event of a tie, both teams get 50 points. Each individual's cost to voting is private information, known right before voting, and distributed uniformly over the interval $[0, 80]$. Thus, each individual's payoff was the team prize (100, 50, or 0) minus the voting cost if they voted or 0 if they abstained. The game was played three times by members of the same group, and subjects received feedback at the end of each round on the number of votes cast by both teams and well as the winning team and their own payoff for the round.

This voter turnout game has a number of testable predictions stemming from the Bayesian Nash equilibrium (hereafter, BNE). First, turnout for both teams should decline with the total group size, so it should be smaller in the Large treatment as compared with the Small treatment. Second, regardless of group size, the turnout rate should be higher for the minority team as compared with the majority team in order to offset the size advantage of the latter group (this is also known as the "underdog effect").

Figure 2 shows mean turnout rates with 95% CIs and BNE predictions across experiments, treatments, and rounds. First, observed turnout is generally much greater than predicted for either team in both the Large and Small group treatments across both experiments. One ex-

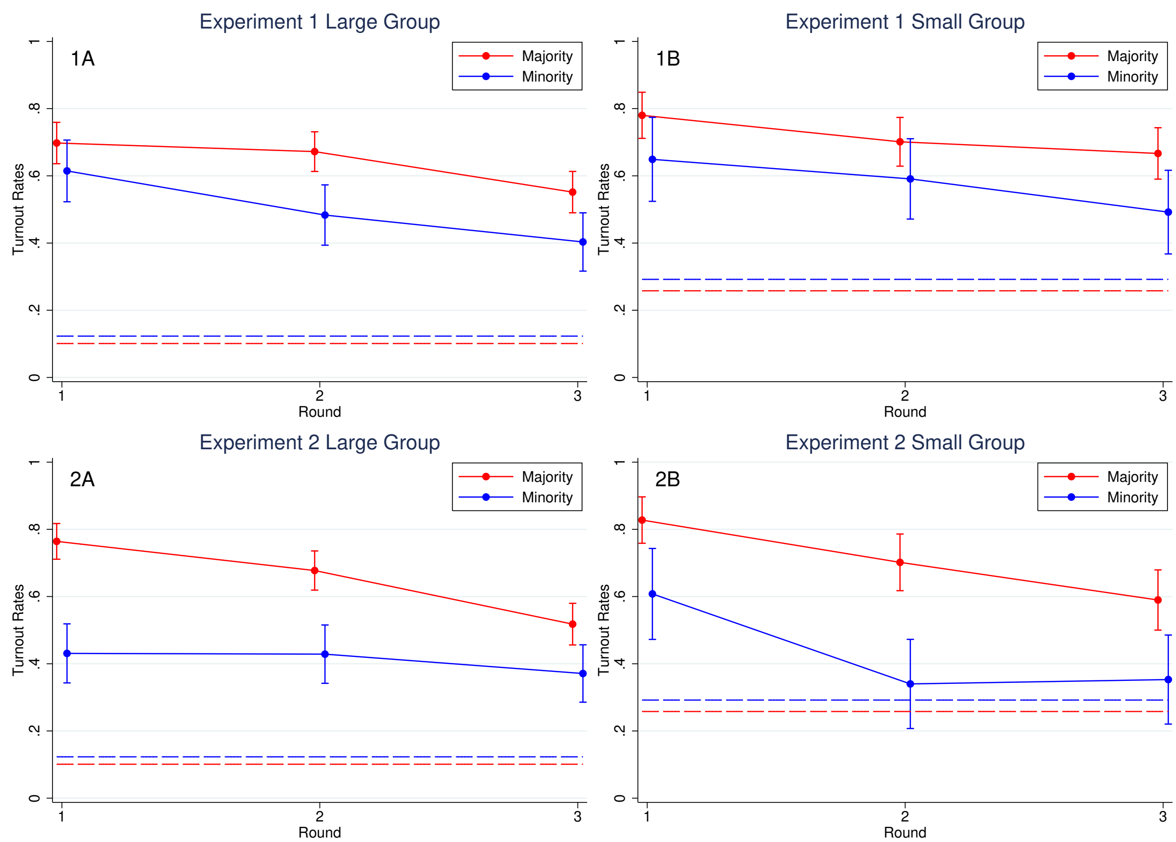

Fig. 2: Majority and minority team turnout rates under different electorate sizes. (1A and 1B) The turnout rates for large groups and small groups from Experiment 1. (2A and 2B) The turnout rates for large groups and small groups from Experiment 2. The red curve is the turnout rate for the majority team (with 95% CI). The blue curve is the turnout rate for the minority team (with 95% CI). The dotted lines are the theoretical predictions of the majority and minority turnout rates.

ception is turnout by the minority team in the Small group treatment of Experiment 2, where, in rounds 2 and 3, we find no statistical difference between the mean and predicted turnout rates for the minority team (two-tailed t-test Round 2: $t(48) = 0.7996$, $p = 0.4279$; Round 3: $t(49) = 0.7093$, $p = 0.4815$).

Second, while in Experiment 1 turnout rates for the *majority* teams are significantly lower by round 3 in Large groups than in Small groups (Mann-Whitney test of differences, p-value = 0.0241), there are no corresponding differences for the *minority* teams across treatments. Similarly, in Experiment 2 we find no difference in turnout rates for majority or minority teams across treatments ($p > 0.100$ for all three rounds), except for minority team members in round 1. See the Supplementary Material B.2 for a detailed analysis.

Third, counter to equilibrium predictions, we *do not* observe an underdog effect; turnout rates for the majority team are always higher in both the Large and Small group treatments than for the minority team. See supplementary Table 7 and 8 for details. This finding is nevertheless consistent with many other experimental team participation/voting game studies under majority rule e.g., (*21–24*).[3]

Linear Public Good Game

Our third game is a linear public good game based on (*25*). In this game, subjects are assigned membership to a group of size $N = 10$ (small) or 100 (large). In each round, each group member, i, is endowed with ω tokens and must decide simultaneously and without communication how many tokens, $0 \leq x_i \leq \omega$ to contribute to a public good. Player i's payoff in points is given by:

$$\pi_i = \omega - x_i + \beta \sum_{j \in N} x_j$$

Our experiment used standard parameterizations from the literature, where $\omega = 20$ and the marginal per capita return (MPCR) $\beta = 0.3$. The game was played repeatedly for 8 or more rounds, but we truncate to 8 rounds for comparison across treatments. Following each round of play, subjects received feedback on the group contribution and their own round payoff.

The dominant strategy Nash equilibrium is that subjects contribute 0 to the public good in all rounds due to the fact that the MPCR $\beta < 1$ so that the the marginal return on investments made to the private account (1) dominates that from the public good (β). However, it is *socially optimal* if all N players contribute their entire endowment, since $N\beta > 1$. These predictions are invariant to group size, though one might expect contributions to be lower in larger groups owing to the greater temptation to free ride on the contributions of others or the greater perceived social pressures in smaller groups to contribute more (*26*).

Figure 3 reports mean contributions to the public good as a percentage of subjects' endowment across experiments, treatments, and rounds. Contributions decrease over time as subjects gain experience. But there is little to no difference in contributions between Large and Small groups over each of the 8 rounds. In neither experiment, we find significant differences in average contributions over all rounds (Mann-Whitney test with pooled data over all 8 rounds, Experiment 1: $p = 0.343$; Experiment 2: $p = 0.405$). These results are generally consistent with the ambiguous effects of group size on public good contributions found in the literature. For instance, a meta-analysis of 27 linear public goods game experiments by (*27*) found insignificant effects of group size for public good contributions.

The patterns of behavioral types show a key difference of contributions between Large and Small groups. In Large groups we find a greater number of extreme behavioral types in "strong" free-riders (Mann-Whitney test with pooled data over all 8 rounds, Experiment 1: $p < 0.001$; Experiment 2: $p < 0.001$) and altruists (Mann-Whitney test with pooled data over all 8 rounds, Experiment 1: $p < 0.001$; Experiment 2: $p = 0.004$). This greater heterogeneity washes out in the aggregate, resulting in similar average contributions in both the Large and Small groups. See Supplementary Material B.3 for definitions and a detailed analysis.

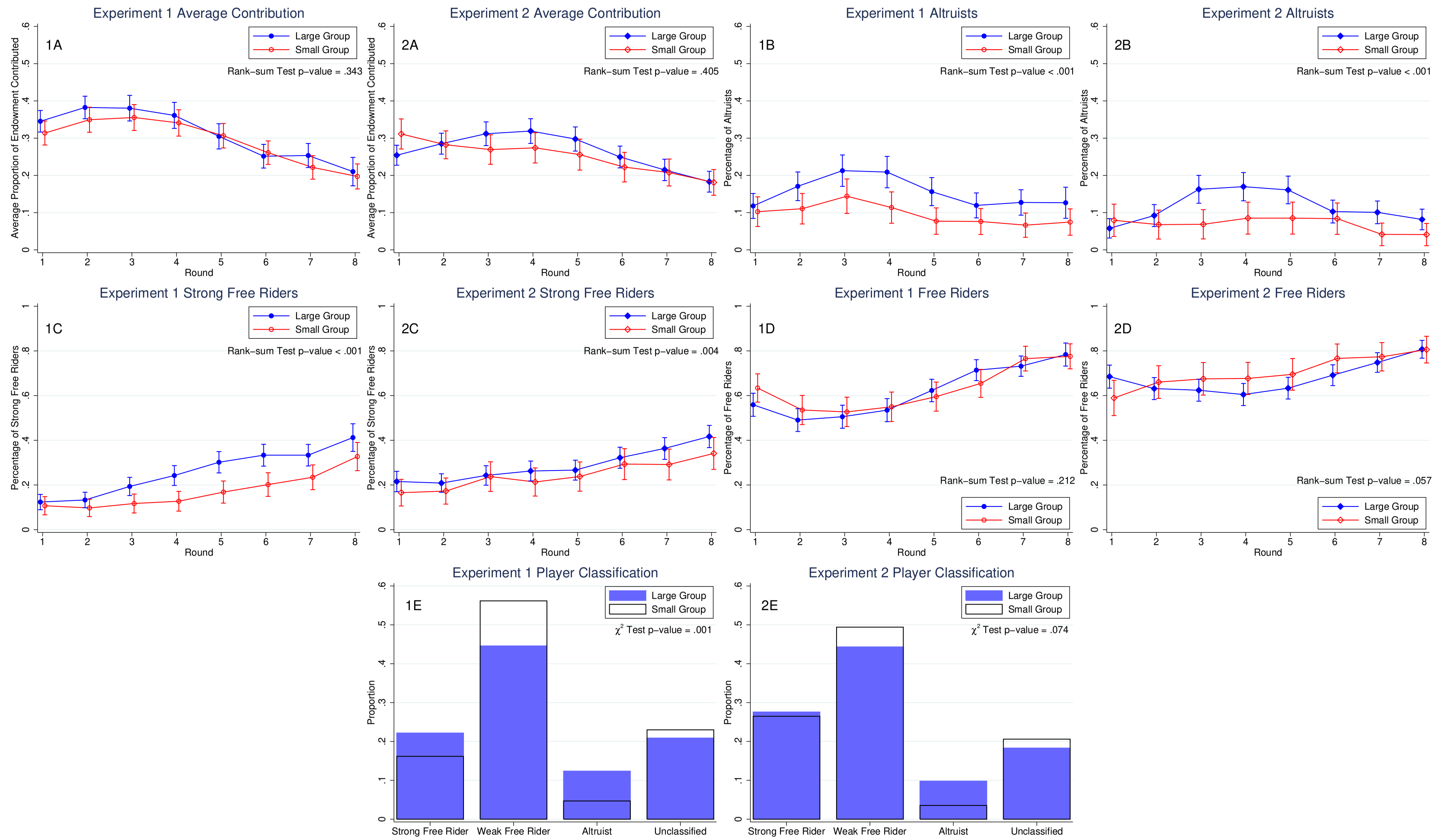

Fig. 3: Average contribution, proportion of (strong) free-riding and altruists and player classifications in the public good games. (1A-1D) The average proportion of endowment contributed, proportion of (strong) free-riding and altruists (overlaid with 95% CI) across different rounds in Experiment 1 with the p-values from a Mann-Whitney test on pooled data over all eight rounds. (2A-2D) The average proportion of endowment contributed, proportion of (strong) free-riding and altruists (overlaid with 95% CI) across different rounds in Experiment 2 with the p-values from a Mann-Whitney test on pooled data over all eight rounds. (1E and 2E) The classification results of Experiment 1 and 2 with the p-values of χ^2 tests.

Behavior Across Games

In both experiments, the first three games were followed by the next five 2-player or individual decision-making games/tasks in the order played: 4) a 2-player ultimatum game played once (in Experiment 2 we reversed roles and played it twice), 5) an individual Holt-Laury type risk preference elicitation played once, 6) a 2-player centipede game played 3 times, 7) a 2-player trust game played once (in Experiment 2 we reversed roles and played it twice), 8) an individual 3-round real effort task exploring gender differences in compensation schemes. Details and results for these games are in Supplementary Material B.

Over the past three decades, behavioral and experimental economists have made huge progress in identifying the stylized facts of these games (see (*28*)). However, an empirical understanding of how strategic behavior varies across these games remains an open question.

The multi-game protocol of our experiment allows us to observe each player's strategic profile across many different games and hence to look for an empirical relationship in strategic behavior between games. If a player follows the Nash Equilibrium, then he might believe his opponents are also rational and best respond to such a belief. Moreover, he has to believe his opponents would also best respond to a rational strategic profile so that the choices and beliefs are consistent in equilibrium, yielding perfect behavioral correlations across games.

Reaching an equilibrium requires that all players behave fully rationally. If this is not the case, players may choose non-equilibrium strategies and correlations in behavior across games are weakened. For instance, (*29*) and (*30*) suggest that observed strategic sophistication are not only determined by the reasoning ability of subjects but also by their beliefs about opponents' sophistication. Psychological factors other than bounded rationality, such as social preferences, can also drive players to deviate from the equilibrium. However, it is not clear if there is any association in dis-equilibrium behavior across different games. Here we take advantage of our design to investigate correlations in strategic measures across games or attributes.

Specifically, we use two approaches to analyze these correlations. First, we explore how an individual's sophistication level, as classified by their first guesses in the p-beauty contest game (also their first choice in the experiment), is correlated with their behavior in the other seven games by a subsample analysis. Second, we compute pairwise Spearman's rank-order correlation coefficients of raw choices in each of the eight games/tasks to summarize the empirical relationship across all games.

We construct 7 measurements of strategic behavior and risk attitudes. Three measurements are related to social preferences: whether a player exhibited strong free-riding behavior in the public good game, the proposal offer made in the ultimatum game and the investment made in the trust game. Three other measurements capture strategic sophistication: the initial guess in the beauty contest game, the average take rate in the first decision node of the centipede game and the frequency of following the BNE in the voter turnout game. The last is the number of safe options chosen in the risk elicitation task, measuring risk aversion. See Supplementary Material B for the analysis of each measurement and Supplementary Material E for additional subsample analysis. Supplementary Table 4 summarizes all 7 behavioral measurements.

Subsample Analysis

We disaggregate the data according to subjects' initial guesses in the p-beauty contest game using a level-k classification. Following (*19*), we assign level 0 to those choosing the midpoint of the guessing interval (or higher). Thus, a level 0 ($L0$) subject would guess 50 or higher. A level 1 ($L1$) subject would guess less than 50, but greater than or equal to $2/3 \times 50 = 33.33$, i.e., in $[33.33, 50)$, and so on. We group subjects with a guess less than 14.81 in the most sophisticated category ($> L3$). Supplementary Figure 4The distribution of levels based on the initial guess in the beauty contest game. The blue bars and red bars show the classification results from Experiments 1 and 2, respectively. The p-value of the Kolmogorov-Smirnov test is provided in the figure.figure.caption.15 shows the classification result, indicating no significant difference between Experiments 1 and 2 (Kolmogorov–Smirnov Test: $KS = 0.0727$, $p = 0.148$).

We report on the means and 95% CIs of other game/task measures disaggregated by level type in Figure 4. Non-parametric trend tests are used to identify monotonicity. For instance, the first panel shows how the risk aversion index varies with the level types, where we find no monotonic trend in either Experiment 1 or 2 (Experiment 1: $p = 0.506$; Experiment 2: $p = 0.119$). Similarly, we find no significant monotonic trends in the average take rate at the first decision node of the centipede game (Experiment 1: $p = 0.221$; Experiment 2: $p = 0.073$) or the amount of investment in the trust game (Experiment 1: $p = 0.204$; Experiment 2: $p = 0.718$).

By contrast, we find significant monotonic trends in the frequency of being a strong free-rider in the public good game (Experiment 1: $p < 0.001$; Experiment 2: $p = 0.004$) and the proposal offer in the ultimatum game (Experiment 1: $p = 0.003$; Experiment 2: $p = 0.024$). The trend in the frequency of following the BNE threshold strategy in the voter turnout game is the only measure that is not consistent between Experiments 1 and 2. In Experiment 1, more sophisticated players by level-k type are more likely to follow the BNE ($p = 0.018$) but the trend is insignificant in Experiment 2 ($p = 0.711$).

Broadly speaking, the monotonic trends are similar across experiments, suggesting that the underlying cause of the correlations is immune to the experimental population. Focusing on the strategic measures with significant trends, we find that the pattern is consistent with the direction of Nash equilibrium—more sophisticated players are more likely to choose the strictly dominant strategy of contributing nothing in the public good game and their offers are closer to the subgame perfect equilibrium in the ultimatum game. This demonstrates the predictive power of the equilibrium while its sensitivity to the beliefs about other players' behavior.

Spearman's Rank-Order Correlation Coefficients

Figure 5 summarizes the pairwise Spearman's rank-order correlation coefficients for the 7 measures used in the prior section. We apply the Bonferroni correction to counteract the problem of multiple comparisons for the significance testing of the correlation coefficients.

Similar to the subsample analysis, we observe a significant negative correlation between the initial guess and strong free-riding behavior (Experiment 1: $\rho = -0.2129$, $p = 0.0003$; Experiment 2: $\rho = -0.1458$, $p = 0.0136$). Moreover, proposers who offer less in the ultimatum game are more likely to be a strong free-riders (Experiment 1: $\rho = -0.1606$, $p = 0.0923$; Experiment 2: $\rho = -0.2076$, $p = 0.0106$). This result suggests that the players tend to adopt similar reasoning processes in the beauty contest game, the public good game and the ultimatum game—which is consistently in the direction of the Nash equilibrium.

Furthermore, Figure 5 shows that the more risk averse players are significantly less likely to be strong free riders (Experiment 1: $\rho = -0.1380$, $p = 0.0261$; Experiment 2: $\rho = -0.1713$, $p = 0.0011$). This significant correlation is unexpected from the perspective of equilibrium since being a strong free-rider is a strictly dominant strategy. That is, a payoff maximizing player should choose this *regardless* of his risk preference. This finding suggests that while Nash equilibrium has predictive power in understanding dis-equilibrium behavior, it cannot be the sole explanation. Alternatively, models with social image concerns can support such an empirical relationship if players believe there is a non-trivial probability that their behavior is "observed"—viz. the "audience effect" (see (*31*)).

Finally, the insignificant correlation between the risk preference measure and the amount of investment in the trust game is consistent with the finding in (*11*). Conceptually, sending money to another player in the first stage of the trust game is an uncertain prospect and hence the correlation between investment and risk aversion is plausible. Yet, the insignificance indicates the players in the trust game do not view the investment as a gamble. Instead, the weak correlation between the investment amount and the take rate in the centipede game that we found in Experiment 2 ($\rho = -0.2472$, $p = 0.0755$) suggests that investment behavior is potentially related to the belief forming ability in a multi-stage game.

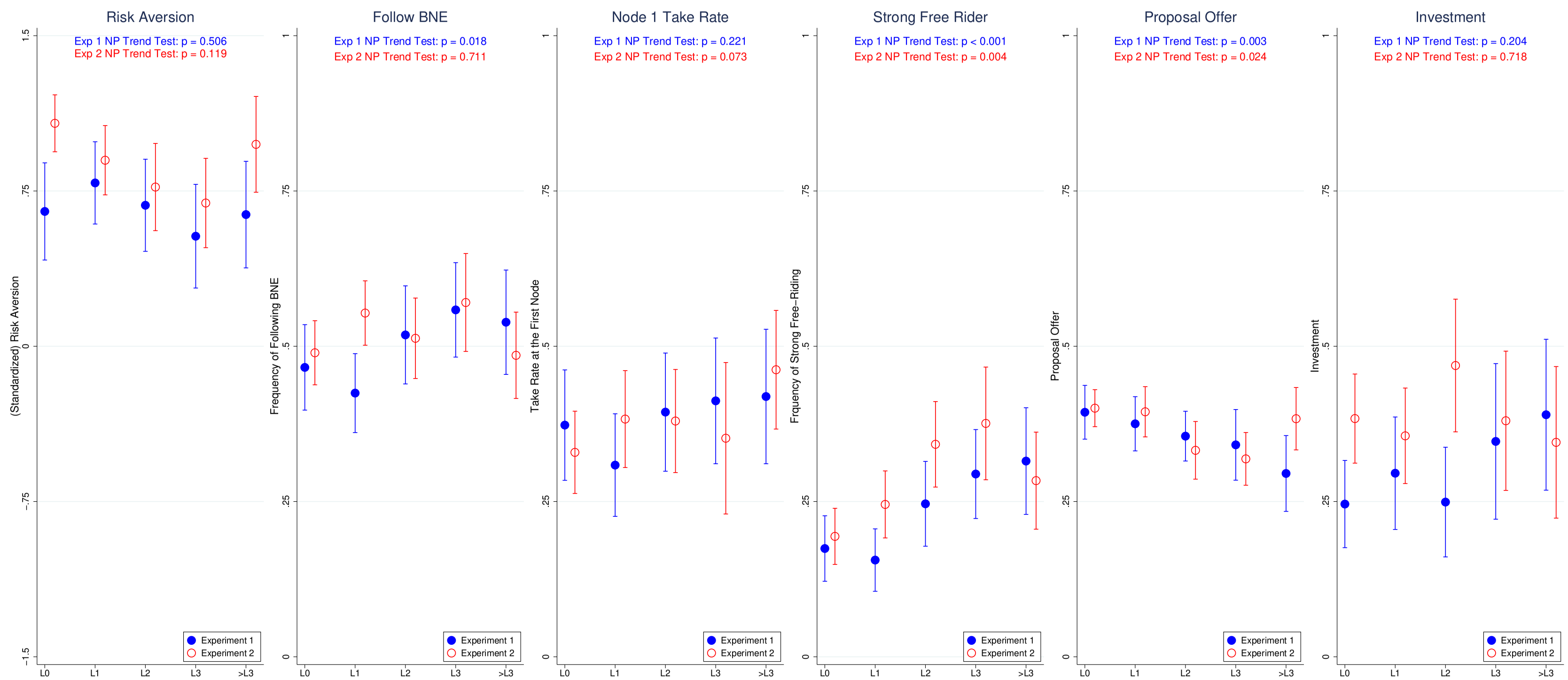

Fig. 4: Means of the game-play behavior for each level classified based on the initial guesses in the beauty contest game. Bars indicate the 95% CIs. The p-values of non-parametric trend tests are provided in the figures.

Experiment 1 Spearman Correlation Matrix

	Risk Aversion	Initial Guess	Follow BNE	Take Rate at Node 1	Strong Free Rider	Proposal Offer	Investment
Risk Aversion	1.0000						

Initial Guess	0.0276	1.0000					
	[1.0000]	---					
Follow BNE	0.0109	-0.1056	1.0000				
	[1.0000]	[0.7108]	---				
Take Rate at Node 1	-0.0232	-0.0979	0.0168	1.0000			
	[1.0000]	[1.0000]	[1.0000]	---			
Strong Free Rider	-0.1380	-0.2129	0.0989	0.1196	1.0000		
	[0.0261]	[0.0003]	[0.3197]	[0.7412]	---		
Proposal Offer	0.0002	0.2046	-0.0644	-0.0437	-0.1606	1.0000	
	[1.0000]	[0.0637]	[1.0000]	[1.0000]	[0.0923]	---	
Investment	-0.0329	-0.0645	0.0436	-0.1642	-0.0571	-0.0465	1.0000
	[1.0000]	[1.0000]	[1.0000]	[0.8931]	[1.0000]	[1.0000]	---

Positively Correlated [p<0.01] [p<0.05] [p<0.10] Uncorrelated Negatively Correlated [p<0.10] [p<0.05] [p<0.01]

Experiment 2 Spearman Correlation Matrix

	Risk Aversion	Initial Guess	Follow BNE	Take Rate at Node 1	Strong Free Rider	Proposal Offer	Investment
Risk Aversion	1.0000						

Initial Guess	0.0610	1.0000					
	[1.0000]	---					
Follow BNE	-0.0051	-0.0163	1.0000				
	[1.0000]	[1.0000]	---				
Take Rate at Node 1	0.0684	-0.1075	0.0902	1.0000			
	[1.0000]	[1.0000]	[1.0000]	---			
Strong Free Rider	-0.1713	-0.1458	0.0893	0.1152	1.0000		
	[0.0011]	[0.0136]	[0.7676]	[1.0000]	---		
Proposal Offer	0.1524	0.1436	-0.0755	0.0442	-0.2076	1.0000	
	[0.2358]	[0.3688]	[1.0000]	[1.0000]	[0.0106]	---	
Investment	-0.1062	0.0048	0.0334	-0.2472	0.0016	0.0547	1.0000
	[1.0000]	[1.0000]	[1.0000]	[0.0755]	[1.0000]	[1.0000]	---

Positively Correlated [p<0.01] [p<0.05] [p<0.10] Uncorrelated Negatively Correlated [p<0.10] [p<0.05] [p<0.01]

Fig. 5: Pairwise Spearman rank-order correlation coefficients of strategic behavior. Color-coded correlation coefficients indicate statistical significance with Bonferroni corrections. We use warm colors for positive correlations and cold colors for negative correlations. A darker color implies the correlation coefficient is more statistically significant. The left panel shows the correlation matrix for Experiment 1 and the right panel is for Experiment 2.

Discussion

We have demonstrated the possibility of conducting large, repeated, multi-game economic experiments with mobile phones and mobile payments. Our results, particularly for small groups and 2-player games or individual decision making tasks are similar to findings from traditional laboratory experiments which provides reassurance that our mobile platform does not appear to affect subject behavior. At the same time, we are able to leverage our large-scale, multi-game approach to obtain interesting new findings on group size effects and correlations in strategic behavior across games that would be difficult to obtain in traditional, limited capacity laboratory settings. We have further demonstrated that our approach does not require that subjects participate *in-person*; we have repeated our experiment a second time with subjects playing the same set of games remotely, and we obtain similar results. This replication of our original findings one year later with similar results builds confidence in the robustness of our findings and suggests that differences in the degree of physical control and/or social context between participating in-person or remotely may be less impactful on subject behavior for the scenarios we tested than was previously thought.

Our methodology provides an exciting and promising way forward for experimental research in the post-COVID-19 era, not only for classic lab-type experimentation but also for randomized control trial (RCT) experiments conducted in the field and remotely online. Indeed, our approach blurs the difference between laboratory and field RCTs.

Experiments hosted on the cloud can scale efficiently to run with hundreds or thousands of in-person or remote participants. User interfaces and incentives can be kept consistent across device types to ensure comparability of results. Combined with mobile payments, our tools and approach can significantly lower the costs of and create new opportunities for running laboratory, field, and online experiments.

We recognize that there are rival, labor-saving ways of implementing the large scale, repeated multi-game experiment that we conducted and report on in this paper. For instance, social science researchers can and have used on-line workers (e.g., on Amazon's mechanical Turk) or large, on-line panels of subjects (e.g., Prime panels). However, using common-pool online workforces also has its drawbacks; worker payments are low and set by market-wide conditions rather than by the experimenter; there is the risk of bot players, or players playing on multiple accounts at the same time leading to more screening of subjects and cleaning of data; finally, players on such platforms might be very experienced (less naive, less pro-social) about social science research questions than is the general population. Our approach allows for the quick recruitment of any sample of subjects (including the traditional sample of university student subjects), with fewer of the downsides of online workforces or panels.

Our findings provide the first important proof of concept for a new methodology of controlled experimentation using mobile platforms and payments. We hope that others build upon our approach.

Method

Our experiments were conducted using MobLab, an online, cloud-based educational platform for conducting economics experiments using web browsers and/or mobile devices. The eight games used in the experiment were pre-programmed by MobLab.[4] In both Experiment 1 and in the remotely conducted Experiment 2, we presented game-specific instruction slides before the start of each game so that subjects had no prior knowledge of subsequent games. The screenshots, instructions, and configurations of the games can be found in Supplementary Material F.2.

There were 633 players in Experiment 1 and 585 players in Experiment 2 who participated in at least one of the eight games. For both Experiments 1 and 2, students at the summer camp had recently completed their third year of university study. On the day before the experiment, a survey was sent to the participants to collect demographic data on gender, place of origin, cognitive reflection test (CRT) scores and participant's (self-reported) score on China's National College Entrance Examination (NCEE), commonly known as the Gaokao score.

Experiment 1 took place in an auditorium with a capacity of 800 seats. Before the experiment, students were randomly assigned to two matching pools, section A (large groups) and section B (small groups), and seated in separate areas of the auditorium. It was made common knowledge that the players would only be matched with other players within the same section (either A or B) through all eight games, and group membership was shuffled between but not within games.

The 2020 Summer Camp and Experiment 2 were held online due to the COVID-19 pandemic. The recruitment procedure was the same as in Experiment 1 while the implementation of the randomization was different. Participants were randomly assigned and invited by email to join two separate MobLab classes. Players would only match against other players within their class / matching group. As in Experiment 1, it is common knowledge to all subjects that they would only be matched with the subjects in the same matching group while being re-matched before each game. See Supplementary Material F.1 for details of the implementation.

At the beginning of both experiments, we communicated to subjects that their decisions and corresponding points earned from all games would matter for their final payment. They were told to expect a show-up payment of 10 CNY for participating in the experiment. The average payoff per game was 3 CNY with a final payment being the sum of participant's show-up payment and total payoff across all 8 games. In Experiments 1 and 2, the overall average total payment was 37.61 CNY ($\approx$ 5.42 USD) and 40.00 CNY ($\approx$ 5.77 USD), respectively, or roughly the equivalent of 2 hours of work as an undergrad TA in China. Subjects were paid on Alipay, the payments platform of Alibaba, which is ubiquitous in China and is also the world's largest mobile payment platform. Their account information was collected before the experiment with consent and was only used for this experiment. Due to technical issues and frustrations faced by participants during some parts of Experiment 1, we decided to increase and smooth the average payment per participant, ex post. For each successfully completed game, a participant was awarded a payoff of 0.0205 CNY $\times$ the points earned in that game. For games that could not be joined or completed, they were awarded the average payoff earned by participants who had successfully completed that game. The final participant payoff was the the sum of their show-up payment of 10 CNY and their game based payment with a minimum payment floor of 27 CNY for all participants.

A weakness of our within-subjects design is that the order of play of the games may matter, for instance there may be spillover effects from one game to the next. However, our ability to replicate many results under somewhat *weaker control* conditions than is typically employed, speaks to the *robustness* of those experimental findings.

Acknowledgements

We thank Colin Camerer, Joseph Tao-yi Wang, Thomas Palfrey, and audiences at the University of Queensland Behavioural and Economics Science Cluster e-seminar for helpful feedback. We also thank the MobLab team for developing and supplying the experiment platform and providing technical assistance both on-site and remotely throughout our experiments. Finally, we thank the student assistants from Xiamen University for helping with testing and logistics. **Funding:** Z.L. was supported by National Natural Science Foundation of China (Grant No. 71873116) and the NSFC Basic Science Center Program (Grant No. 71988101). Experiment 1 was funded by MobLab. Experiment 2 was funded by Xiamen University. **Author contributions:** Z.L., P.L., S.K., D.W. and J.D. designed the research. S.K. managed the development and testing of the experiment software. Z.L., S.K. and D.W. conducted the experiments. Z.L. and

P.L. analysed the data. Z.L., P.L., S.K., D.W. and J.D. wrote the paper. All authors discussed the results and implications and commented on the manuscript. **Competing interest:** Z.L. has no competing interest. P.L. was employed by MobLab from December 2017 to December 2018 and compensated more than US$10,000 during the last 3 years. S.K. and D.W. are employed by MobLab and compensated more than US$10,000 during the last 3 years. J.D. is a Scientific Advisor to MobLab, a position with no compensation but with a small equity stake. **Data and materials availability:** The experimental data and code for all analyses will be available on Open Science Framework once the manuscript is accepted.

Notes

[1]Online workforces such as Amazon's Mechanical Turk provide a similar opportunity for large-scale experimentation, but they cannot be conducted *in-person* and they involve less control over participants' attention, participation, and interactions, especially in some of the multi-player games that we study here. See (*32*) for a further discussion.

[2]Other experimental studies focusing on group size effects include (*33*) and (*34*) who look at how large groups of players, up to size 100, play the volunteer's dilemma; (*35*), (*36*) and (*37*) who look at asset pricing in markets with large number of subjects, between 40-300.

[3]As (*24*) point out (footnote 4), the *only* experimental paper reporting greater turnout by minority team members in a majority rule setting is (*9*).

[4]In July 2019, we were alerted to an opportunity to conduct an experiment on a group of college students participating in a summer camp at Xiamen University. We had only about 20 days advance notice to design and implement an experiment that could make use of the population of more than 600 students attending the camp. For this opportunity, we selected from among a pre-programmed suite of economic experiments developed by MobLab that could run on mobile devices, scale well to large groups, and offer potentially interesting cross-game correlations in behavior and outcomes to study. The design and implementation of the experiment was reviewed and approved by the organizing committee of the summer camp from the School of Economics, and the Managing Department (Division) of Social Sciences at Xiamen University.

References

1. W. Güth, R. Schmittberger, B. Schwarze, *Journal of Economic Behavior & Organization* **3**, 367 (1982).

2. C. A. Holt, S. K. Laury, *American Economic Review* **92**, 1644 (2002).

3. R. D. McKelvey, T. R. Palfrey, *Econometrica* pp. 803–836 (1992).

4. J. Berg, J. Dickhaut, K. McCabe, *Games and Economic Behavior* **10**, 122 (1995).

5. M. Niederle, L. Vesterlund, *The Quarterly Journal of Economics* **122**, 1067 (2007).

6. R. M. Isaac, J. M. Walker, A. W. Williams, *Journal of Public Economics* **54**, 1 (1994).

7. D. Nosenzo, S. Quercia, M. Sefton, *Experimental Economics* **18**, 4 (2015).

8. A. Bosch-Domènech, J. G. Montalvo, R. Nagel, A. Satorra, *American Economic Review* **92**, 1687 (2002).

9. D. K. Levine, T. R. Palfrey, *American Political Science Review* **101**, 143 (2007).

10. M. Faravelli, K. Kalayci, C. Pimienta, *Experimental Economics* pp. 1–25 (2019).

11. M. Dean, P. Ortoleva, *Proceedings of the National Academy of Sciences* **116**, 16262 (2019).

12. J. Chapman, M. Dean, P. Ortoleva, E. Snowberg, C. Camerer, Econographics, *Tech. rep.*, National Bureau of Economic Research (2018).

13. C. F. Camerer, T.-H. Ho, J.-K. Chong, *The Quarterly Journal of Economics* **119**, 861 (2004).

14. T. C. Burnham, D. Cesarini, M. Johannesson, P. Lichtenstein, B. Wallace, *Journal of Economic Behavior & Organization* **72**, 171 (2009).

15. D. Gill, V. L. Prowse, *Available at SSRN 2043336* (2012).

16. J.-K. Chong, C. F. Camerer, T.-H. Ho, *Experimental Business Research* (Springer, 2005), pp. 203–228.

17. S. Georganas, P. J. Healy, R. A. Weber, *Journal of Economic Theory* **159**, 369 (2015).

18. F. Cerigioni, F. Germano, P. Rey-Biel, P. Zuazo-Garin, *et al.*, Higher orders of rationality and the structure of games, *Tech. rep.* (2019).

19. R. Nagel, *The American Economic Review* **85**, 1313 (1995).

20. T.-H. Ho, C. Camerer, K. Weigelt, *The American Economic Review* **88**, 947 (1998).

21. J. Duffy, M. Tavits, *American Journal of Political Science* **52**, 603–618 (2008).

22. J. Großer, A. Schram, *American Journal of Political Science* **54**, 700– (2010).

23. M. Kartal, *Experimental Economics* **18**, 366– (2015).

24. M. Agranov, J. K. Goeree, J. Romero, L. Yariv, *Journal of the European Economic Association* **16**, 825 (2018).

25. R. M. Isaac, J. M. Walker, *The Quarterly Journal of Economics* **103**, 179 (1988).

26. M. Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press, 1965).

27. J. Zelmer, *Experimental Economics* **6**, 299–310 (2003).

28. C. F. Camerer, *Behavioral game theory: Experiments in strategic interaction* (Princeton University Press, 2003).

29. M. Agranov, E. Potamites, A. Schotter, C. Tergiman, *Games and Economic Behavior* **75**, 449 (2012).

30. L. Alaoui, A. Penta, *The Review of Economic Studies* **83**, 1297 (2016).

31. J. Andreoni, B. D. Bernheim, *Econometrica* **77**, 1607 (2009).

32. A. A. Arechar, S. Gächter, L. Molleman, *Experimental Economics* **21**, 99 (2018).

33. A. Franzen, *Rationality and Society* **7**, 183 (1995).

34. A. Kopányi-Peuker, *Journal of Behavioral and Experimental Economics* **80**, 211 (2019).

35. A. W. Williams, J. M. Walker, *The Journal of Economic Education* **24**, 291 (1993).

36. P. Bossaerts, C. Plott, *Review of Finance* **8**, 135 (2004).

37. C. Hommes, A. Kopányi-Peuker, J. Sonnemans, *Experimental Economics* pp. 1–20 (2020).

List of Supplementary Materials

Figures S1–S26

Tables S1–S24